

# Sharp oracle inequalities and slope heuristic for specification probabilities estimation in general random fields

Matthieu Lerasle\* and Daniel Yasumasa Takahashi†

June 14, 2011

## Abstract

We provide new methods for estimation of the one-point specification probabilities in general discrete random fields. Our procedures are based on model selection by minimization of a penalized empirical criterion. The selected estimators satisfy sharp oracle inequalities without any assumption on the random field for both  $L_2$ -risk and Kullback loss. We also prove the validity of slope heuristic for the specification probabilities estimation problem. We finally show in simulation studies the practical performances of our methods.

## 1 Introduction

Random fields are used in variety of domains including computer vision [Bes93, Woo78], image processing [CJ83], neuroscience [SBSB06], and as a general model in spatial statistics [Rip81]. The main motivation for our work comes from neuroscience where the advancement of multi-channel and optical technology enabled the scientists to study not only a unit of neurons per time, but tens to thousands of neurons simultaneously [TSMI10]. The important question in neuroscience is to understand how the neurons in this ensemble interact with each other and how this is related to the animal behavior [SBSB06, BKM04]. This question turns out to be hard for three reasons. First, the experimenter has always only access to a small part of the neural system, which means that the system is partially observed. Also, there is no good and tractable model for population of neurons in spite of the good models available for single neurons, therefore very general models must be considered. Finally, strong long range multi-neuron interactions exist [LOU<sup>+</sup>10]. Our work overcomes these difficulties as will be shown.

A random field is a triplet  $(S, A, P)$  where  $S$  is a discrete set of *sites*,  $A$  is a finite alphabet and  $P$  is a probability measure on the set  $\mathcal{X}(S) = A^S$  of *configurations* on  $S$ . Given a random field  $(S, A, P)$ , we define the *one point specification probabilities* of  $P$  as regular versions of the following conditional probabilities, for all sites  $i$  in  $S$ , for all configurations  $x$  in  $\mathcal{X}(S)$ ,

$$P_{i|S}(x) = P(x(i)|x(j), j \in S/\{i\}).$$

The specification probabilities are important in the applications as they encode some conditional independence between the sites, see for example [BM09, BMS08, CT06a, GOT10, RWL10, LT11]. The main goal of this paper is to provide good estimators of the specification probabilities. We do not assume that the set of sites  $S$  is finite. However, the set of observed sites,

\*Instituto de Matemática e Estatística; Universidade de São Paulo; Caixa Postal 66281; 05315-970 São Paulo, Brasil; lerasle@gmail.com

†Instituto de Matemática e Estatística; Universidade de São Paulo; Caixa Postal 66281; 05315-970 São Paulo, Brasil; currently at Princeton University; Department of Psychology and Neuroscience; Green Hall, Princeton, NJ 08540; takahashi@d@gmail.com

$V_M \subset S$  is finite, with cardinality  $M$ . Let  $X_{1:n} = (X_1, \dots, X_n)$  be an i.i.d. sample with marginal law  $P$ , the observation set consists in  $X_{1:n}(V_M) = (X_i(j))_{i=1, \dots, n; j \in V_M}$ . As all the results are non asymptotic,  $M$  is allowed to grow with  $n$ .

This model enables us to handle the following situations that are of particular importance in neuroscience.

**Example 1: Dynamic estimation of the connected neurons**

$S$  is composed by  $T$  copies of the set of all neurons,  $V_M$  is composed by  $T$  copies of the set of observed neurons. A configuration represents the neural activity in a period of time of size  $T$ . For all neurons  $i$  at time  $t'$ , the support of  $P_{i,t'|S}$  defined by the minimal subsets  $C$  of  $S$  such that  $P_{i,t'|S} = P_{i,t'|C}$  represents the set of neurons connected to  $i$  in a period of time of size  $T$ . In practice  $C$  usually is not totally contained in  $V_M$  and we don't know the shape of  $P$ . The problem is to obtain a good approximation of  $C$ , observing only the configurations in  $V_M$ .

**Example 2: Prediction of animal behavior**

$S$  is composed by the union of  $T$  copies of the set of all neurons and  $T$  copies of an other site  $\mathcal{I}_o$ ;  $V_M$  is composed by the union of  $T$  copies of the set of observed neurons and  $T$  copies of  $\mathcal{I}_o$ . A configuration represents the neural activity and an associated animal behavior response  $(x(\mathcal{I}_o, t))_t$  in a period of time of size  $T$ . The support  $C$  of  $P_{\mathcal{I}_o, t|S}$  represents then the set of neurons that should be observed during a period of time smaller than  $T$  around  $t$  to predict the behavior of the animal at time  $t$ . Again, the problem is to obtain a good approximation to  $C$  knowing very few about the animal behavior and the neural system.

Our estimators are derived from a model selection procedure by minimization of a penalized empirical criterion. This procedure selects a subset  $\hat{V}$  with cardinality  $|\hat{V}| = O(\ln n)$ . Our first result is that the empirical conditional probabilities  $\hat{P}_{i|\hat{V}}$  as estimators of  $P_{i|S}$  satisfy a sharp oracle inequality (see Section 2 and Theorems 3.2 and 3.4 for details).

The second result of the paper is a proof of the slope heuristic for the estimation of specification probabilities. The heuristic, introduced in [BM07], is a data driven way to optimize the constant in front of the penalty term of the selection procedure. This heuristic is very important in practice, because the constants involved in Theorems 3.2 and 3.4 are generally pessimistics (see for example Figure 4 in our simulation study in Section 5).

In most of the applications, the support of  $P_{i|S}$ , defined as the minimal subset  $V_\star \subset S$  such that  $P_{i|V_\star} = P_{i|S}$  (see Section 2 for details) is usually the object of interest. This is why most of the literature focus on the estimation of  $V_\star$  see [BM09, BMS08, CT06a, GOT10, RWL10] for example. This approach requires in general strong assumptions on the random field, e.g., to be Ising models with strong conditions on the temperature parameter [BM09, GOT10, RWL10]. In particular, [BM09, BMS08, RWL10] assumed that the set  $S$  is finite and that all the sites are observed, i.e that  $V_M = S$ . When  $V_M$  does not contain  $V_\star$ , the meaning of the estimator in these paper is not clear. [CT06a] considered  $S = \mathbb{Z}^d$  but assumed that  $V_\star$  is finite. Finally, [GOT10, LT11] worked with infinite sets of sites and without a priori bound on the number of interacting sites but required a two-letters alphabet  $A$  and some assumptions on  $P$  that the practitioner can not easily verify. These restrictions are severe for applications, e.g., in neuroscience, and cast doubt on the theoretical support for application of these methods in practice. Our model selection procedure does not suffer from these drawbacks.

We focus here on the estimation of the conditional probabilities and we develop the oracle approach introduced in [LT11]. As we already noticed in this paper, an oracle  $\hat{V}$  provides a nice estimator of the support of  $P_{i|S}$ . In [LT11], we used the  $L_\infty$ -norm to measure the risk of the estimators. We use now the  $L_2$ -norm and the K ullback loss and the new results do not require any restriction on the random field. In particular, the finite alphabet  $A$  is not restricted to have two letters,  $P$  does not need to be a Gibbs measure, therefore doesn't need to be a Ising or Potts model, and the size of the support  $V_\star$  of  $P_{i|S}$  can be infinite. To our knowledge this is the

first work with this degree of generality.

Theoretical support for the slope heuristic is currently an active area of research and it has been proven only for very few specific models [BM07, AM09, Ler11b, Ler11a, AB10]. Here, we prove the validity of the slope heuristic for conditional probabilities estimation for  $L_2$  and Küllback loss. Our proof technique is novel and sheds new light on the slope heuristic.

The paper is organized as follows. Section 2 presents the framework and some notations that we use all along the paper. Section 3 gives the model selection procedures and the oracle inequalities satisfied by the selected estimators. Section 4 is devoted to the slope heuristic. We recall the heuristic and state the theorems that justify it in our problem. Section 5 illustrates the results of previous sections using some simulation experiments and Section 6 discuss the results, making a detailed comparison with other works on similar problems. The proofs of the main theorems are postponed to Section 7 and the probabilistic tools used in the main proofs are proved in Section 8.

## 2 Preliminaries

Hereafter, we call random field a triplet  $(S, A, P)$  constituted by a discrete set  $S$  of sites, a finite alphabet  $A$  of spins, with cardinality  $a$  and a probability measure  $P$  on the set of configurations  $\mathcal{X}(S) = A^S$ . More generally, for all subsets  $V$  of  $S$ , let  $\mathcal{X}(V) = A^V$  be the set of configurations on  $V$ . For all  $x$  in  $\mathcal{X}(S)$  and all subsets  $V$  of  $S$ , we denote by  $x(V) = (x(j))_{j \in V}$ . For all  $i$  in  $S$ , for all subsets  $V$  of  $S$ , for all  $x$  in  $\mathcal{X}(S)$  and for all probability measures  $Q$  on  $\mathcal{X}(V \cup \{i\})$ , let

$$Q_{i|V}(x) = Q(x(i)|x(V/\{i\}))$$

be a regular version of the conditional probability. Hereafter, we use the convention that, if  $V$  is a finite subset of  $S$ , if  $Q$  is a probability measure on  $\mathcal{X}(V)$  and  $x$  is configuration in  $\mathcal{X}(V \cup \{i\})$  such that  $Q(x(V/\{i\})) = 0$ , then  $Q_{i|V}$  is the uniform law on  $A$ .

For all probability measures  $Q$  on  $\mathcal{X}(S)$  and for all real valued functions  $f$  defined on  $\mathcal{X}(S)$ . We define the  $L_{2,Q}$ -norm of  $f$  by

$$\|f\|_Q = \sqrt{\int f^2(x) \frac{dQ(x(S/\{i\}))}{a}}.$$

We also define the logarithmic loss of a non-negative function  $f$  defined on  $\mathcal{X}(S)$  by

$$L_Q(f) = \int \ln \left( \frac{1}{f(x)} \right) dQ(x).$$

Let  $P$  be a probability measure on  $\mathcal{X}(S)$  and let  $X_1, \dots, X_n$  be i.i.d  $P$ . We introduce the empirical probability measures  $\hat{P}$  defined on  $\mathcal{X}(S)$  by

$$\hat{P}(x) = \frac{1}{n} \sum_{k=1}^n 1_{X_k=x}.$$

For all subsets  $V$  of  $S$ ,  $\hat{P}_{i|V}$  is an estimator of  $P_{i|S}$ . We define the  $L_2$ -risk of the  $\hat{P}_{i|V}$  by  $\|\hat{P}_{i|V} - P_{i|S}\|_P$ . This risk is decomposed via Pythagoras relation to (see Proposition 8.17)

$$\|\hat{P}_{i|V} - P_{i|S}\|_P^2 = \|\hat{P}_{i|V} - P_{i|V}\|_P^2 + \|P_{i|V} - P_{i|S}\|_P^2.$$

The random term of the risk,  $\|\widehat{P}_{i|V} - P_{i|V}\|_P^2$  is called the variance term, and the deterministic term  $\|P_{i|V} - P_{i|S}\|_P^2$  is called the bias term of the risk.

We also define the K ullback loss of the estimator  $\widehat{P}_{i|V}$  by

$$K(P_{i|S}, \widehat{P}_{i|V}) = L_P(\widehat{P}_{i|V}) - L_P(P_{i|S}).$$

The K ullback risk is also decomposed in a variance term and a bias term thanks to the relation

$$\begin{aligned} K(P_{i|S}, \widehat{P}_{i|V}) &= \left( L_P(\widehat{P}_{i|V}) - L_P(P_{i|V}) \right) + \left( L_P(P_{i|V}) - L_P(P_{i|S}) \right) \\ &= \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)} \right) + \int dP(x(S)) \ln \left( \frac{P_{i|S}(x)}{P_{i|V}(x)} \right) \\ &= K(P_{i|V}, \widehat{P}_{i|V}) + K(P_{i|S}, P_{i|V}). \end{aligned}$$

Let  $V_M$  be a finite subset of  $S$  with cardinality  $M \geq e$  of observed sites and let  $X_{1:n}(V_M) = (X_1(j), \dots, X_n(j))_{j \in V_M}$  be the observation set. For all  $V \subset V_M$ , let  $v = \text{Card}(V)$ . Let  $s > e$  be an integer and let

$$\mathcal{V}_s = \{ V \subset V_M, v \leq s \}, \quad N_s = \text{Card}(\mathcal{V}_s).$$

Let  $\Lambda \geq 100$ ,  $\delta > 1$  and let

$$\mathcal{V}_{s,\Lambda} = \left\{ V \in \mathcal{V}_s, \forall x \in \mathcal{X}(S), P(x(V)) = 0 \text{ or } \widehat{P}(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n} \right\}. \quad (1)$$

$$\mathcal{V}_{s,\Lambda}^{(2)} = \left\{ V \in \mathcal{V}_s, \forall x \in \mathcal{X}(S), P(x(V)) = 0 \text{ or } P(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n} \right\}. \quad (2)$$

Let  $p_* \geq 0$ , and let

$$\mathcal{V}_{s,\Lambda,p_*} = \left\{ V \in \mathcal{V}_{s,\Lambda}, \forall x \in \mathcal{X}(S), P_{i|V}(x) = 0 \text{ or } \widehat{P}_{i|V}(x) \geq p_* \right\}. \quad (3)$$

$$\mathcal{V}_{s,\Lambda,p_*}^{(2)} = \left\{ V \in \mathcal{V}_{s,\Lambda}^{(2)}, \forall x \in \mathcal{X}(S), P_{i|V}(x) = 0 \text{ or } P_{i|V}(x) \geq p_* \right\}. \quad (4)$$

The idea of the sets  $\mathcal{V}_{s,\Lambda,p_*}$  is that we restrict the collections of sets  $V$  to those where the possible configurations are sufficiently observed. This restriction will only be required when we will work with the K ullback loss. The main advantage of the sets  $\mathcal{V}_{s,\Lambda,p_*}$  is that the conditions can be verified in practice. In order to illustrate why we introduced  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ , let us give the following weak Gibbs assumption.

**GA** *There exists  $p_* > 0$  such that, for all finite subsets  $V$ , for all sites  $i$ , and for all  $x$  in  $\mathcal{X}(V)$ ,*

$$P(x(V)) = 0 \text{ or } P_{i|V}(x) \geq p_*.$$

We have (see [Mas07] Proposition 2.5 p20)

$$N_s = \sum_{k=0}^s C_M^k \leq \left( \frac{eM}{s} \right)^s \leq M^s.$$

Hence,  $\ln(a^s N_s \delta) \leq s(\ln(aM)) + \ln \delta$ . We have

$$\frac{nP(x(V))}{\Lambda \ln(2a^s N_s \delta)} \geq \frac{e^{\ln n - s \ln(p_*^{-1})}}{\Lambda(s \ln(aM) + \ln(\delta))} \rightarrow +\infty,$$

if  $(\ln n)^{-1}s < s_* = (\ln p_*^{-1})^{-1}$  and  $\Lambda \ln(M\delta) = O(n^\alpha)$ , where  $\alpha \leq \alpha_* = 1 - s_*$ . In that case, for all  $n \geq n(p_*)$ ,  $\mathcal{V}_s = \mathcal{V}_{s,\Lambda}^{(2)} = \mathcal{V}_{s,\Lambda,p_*}^{(2)}$ .

### 3 Model Selection Results

#### 3.1 The quadratic loss

Our first theorem is a concentration inequality for the variance term of the  $L_2$  risk.

**Theorem 3.1.** *Let  $(S, A, P)$  be a random field and let  $V$  in  $\mathcal{V}_s$ . Then, for all  $\delta > 1$  and all  $0 < \eta \leq 1$ , we have, with probability larger than  $1 - \delta^{-1}$ , each of the following*

$$\left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \leq \frac{6}{a} \left( (1 + 8\eta) \frac{a^v}{n} + \frac{4 \ln(2\delta)}{\eta n} + \frac{9 \ln(2\delta)^2}{\eta^4 n} \right), \quad (5)$$

$$\left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \leq \frac{6}{a} \left( (1 + 8\eta) \frac{a^v}{n} + \frac{4 \ln(2\delta)}{\eta n} + \frac{9 \ln(2\delta)^2}{\eta^4 n} \right). \quad (6)$$

**Comments:**

- The risk of the estimator is upper bounded, with probability larger than  $1 - \delta^{-1}$ , by

$$\left\| \hat{P}_{i|V} - P_{i|S} \right\|_P^2 = \left\| P_{i|V} - P_{i|S} \right\|_P^2 + C(a) \frac{a^v}{n} + C'(a) \frac{(\ln(\delta))^2}{n}.$$

This control depends on the approximation properties of  $V$  through the bias  $\left\| P_{i|V} - P_{i|S} \right\|_P^2$  and on the complexity  $a^v$  of  $V$ . In practice, we would like to find a model  $V$  that optimizes this bound, even though the bias term is completely unknown. This is precisely the aim of the following result.

**Theorem 3.2.** *Let  $(S, A, P)$  be a random field. Let  $K > 1$  and let*

$$\hat{V} = \arg \min_{V \in \mathcal{V}_s} \left\{ - \left\| \hat{P}_{i|V} \right\|_{\hat{P}}^2 + \text{pen}(V) \right\}, \text{ where } \text{pen}(V) \geq \frac{6K}{a} \frac{a^v}{n}.$$

*Then, there exists a constant  $\kappa = \kappa(a, K)$  such that for all  $\delta > 1$ , with probability larger than  $1 - 4\delta^{-1}$ ,*

$$\left\| P_{i|S} - \hat{P}_{i|\hat{V}} \right\|_P^2 \leq \kappa \left( \inf_{V \in \mathcal{V}_s} \left\{ \left\| P_{i|S} - P_{i|V} \right\|_P^2 + \text{pen}(V) \right\} + \frac{(\ln(N_s^2 \delta))^2}{n} \right). \quad (7)$$

*Moreover, when  $K > 2$ , there exists a constant  $\kappa = \kappa(a, K)$  such that, with probability larger than  $1 - 4\delta^{-1}$ ,*

$$\left\| P_{i|S} - \hat{P}_{i|\hat{V}} \right\|_P^2 \leq \left( 1 + \frac{8}{\ln(\delta)} \right) \inf_{V \in \mathcal{V}_s} \left\{ \left\| P_{i|S} - P_{i|V} \right\|_P^2 + \text{pen}(V) \right\} + \kappa \frac{(\ln(N_s^2 \delta))^2}{n}. \quad (8)$$

**Comments:**

- We have  $\ln(N_s^2 \delta) \leq 2s(\ln(M)) + \ln \delta$ . Denoting  $\Gamma_{s,M}(\delta) = 2s(\ln(M)) + \ln \delta$ , with probability larger than  $1 - 4\delta^{-1}$ ,

$$\left\| P_{i|S} - \hat{P}_{i|\hat{V}} \right\|_P^2 \leq \left( 1 + \frac{8}{\ln(\delta)} \right) \inf_{V \in \mathcal{V}_s} \left\{ \left\| P_{i|S} - P_{i|V} \right\|_P^2 + C_1 \frac{a^v}{n} \right\} + C_2 \frac{\Gamma_{s,M}^2(\delta)}{n}.$$

We have found a model that optimizes the bound given by Theorem 3.1, up to the  $s \ln(M)$  term, among all the subsets of  $V$ . Remark that this is only the price to pay to make the bound of Theorem 3.1 uniform over all the subsets of  $\mathcal{V}_s$ .

- A very interesting feature of this result in view of the applications is that it holds without restrictions on the random field  $(S, A, P)$ .

### 3.2 The Küllback Loss

Our first result is a sharp control of the variance term of the Küllback risk.

**Theorem 3.3.** *Let  $(S, A, P)$  be a random field, let  $\Lambda \geq 100$ ,  $\delta > 1$ ,  $s > 0$ . Let  $\mathcal{V}_{s,\Lambda}$  be the collection defined in (1). Then, with probability larger than  $1 - \delta^{-1}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda}$ , for all  $\eta > 0$ , we have*

$$\begin{aligned} \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\hat{P}_{i|V}(x)} \right) &\leq 5 \left( (1 + \eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \\ \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right) &\leq 4 \left( (1 + \eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \end{aligned}$$

Let  $\mathcal{V}_{s,\Lambda}^{(2)}$  be the collection defined in (2). Then, with probability larger than  $1 - \delta^{-1}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda}^{(2)}$ , for all  $\eta > 0$ , we have

$$\begin{aligned} \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\hat{P}_{i|V}(x)} \right) &\leq 5 \left( (1 + \eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \\ \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right) &\leq 4 \left( (1 + \eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \end{aligned}$$

#### Comments:

- The variance part of the Küllback risk is controlled as the variance part of the  $L_2$  risk. We only have to restrict the study to the subset  $\mathcal{V}_{s,\Lambda}$  of  $\mathcal{V}_s$  where all the possible configurations are sufficiently observed. This restriction is not important when  $s \ll n$ , and our result holds also without restriction on the random field.

As in the previous section, we want to optimize the bound on the Küllback loss given by Theorem 3.3 among  $\mathcal{V}_{s,\Lambda}$ . We introduce for this purpose the following penalized estimators.

$$\hat{V} = \arg \min_{V \in \mathcal{V}_{s,\Lambda,p_*}} \left\{ - \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \hat{P}_{i|V}(x) \right) + \text{pen}(V) \right\}. \quad (9)$$

$$\hat{V}_{(2)} = \arg \min_{V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}} \left\{ - \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \hat{P}_{i|V}(x) \right) + \text{pen}(V) \right\}. \quad (10)$$

The following theorem shows the oracle properties of the selected estimator when the penalty term is suitably chosen.

**Theorem 3.4.** *Let  $s > 0$ ,  $\delta > 1$ ,  $p_* > 0$ ,  $\Lambda \geq 100$  and let  $\mathcal{V}_{s,\Lambda,p_*}$  and  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$  be the collections defined in (3) and (4). Let  $K > 1$  and let  $\hat{V}$  and  $\hat{V}_{(2)}$  be the penalized estimators defined in (9) and (10) with*

$$\text{pen}(V) \geq 9K \frac{a^v}{n}.$$

Then, we have, for all  $\eta > 0$ , with probability larger than  $1 - 3\delta^{-1}$ ,

$$\frac{1 - \eta}{1 + \eta} K_P(P_{i|S}, \hat{P}_{i|\hat{V}}) \leq \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}} \left\{ K_P(P_{i|S}, P_{i|V}) + \text{pen}(V) \right\} + \left( 2 \ln n + \frac{C_{\Lambda,p_*,K}}{\eta} \right) \frac{\ln(N_s^2 \delta)}{n}.$$

Also, for all  $\eta > 0$ , with probability larger than  $1 - 3\delta^{-1}$ ,

$$\frac{1-\eta}{1+\eta} K_P(P_{i|S}, \hat{P}_{i|\hat{V}_{(2)}}) \leq \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}} \{ K_P(P_{i|S}, P_{i|V}) + \text{pen}(V) \} + \left( 2 \ln n + \frac{C_{\Lambda,p_*,K}}{\eta} \right) \frac{\ln(N_s^2 \delta)}{n}.$$

#### Comments:

- We use the same kind of penalty as in the  $L_2$  case. This is not surprising because the variance parts of the risks were controlled in the same way.
- We do not optimize the bound obtained in Theorem 3.3 among all the sets in  $\mathcal{V}_{s,\Lambda}$ . We have to restrict ourselves to  $\mathcal{V}_{s,\Lambda,p_*}$ . However, the constant  $C_{\Lambda,p_*,K}$  has the form  $p_*^{-1} C_{\Lambda,K}$ . Therefore, we can choose  $p_* = (\ln n)^{-1}$  and optimize the result asymptotically.
- We optimize the bound among all  $\mathcal{V}_s$  under the weak Gibbs assumption **GA**.

## 4 The slope heuristic

In practice, the constants  $6Ka^{-1}$  in Theorem 3.2 and  $9K$  in Theorem 3.4 are a bit pessimistic. In order to optimize these constants, Birgé and Massart [BM07] have introduced the slope heuristic. It states that there is a minimal penalty  $\text{pen}_{\min}$  satisfying the following properties.

- SH1 When  $\text{pen}(V) < \text{pen}_{\min}(V)$ , the complexity of the selected model is as large as possible.
- SH2 When  $\text{pen}(V)$  is slightly larger than the minimal penalty, the complexity of the selected model is much smaller.
- SH3 When  $\text{pen}(V)$  is equal to 2 times the minimal penalty, then the risk of the selected model is asymptotically the one of an oracle.

In practice, the heuristic is used to calibrate the constant in front of the penalty. Suppose that some quantity  $\Delta_V$  proportional to the complexity is known (in the simulations, we will use  $\Delta_V = a^v/n$ , even though we only know thanks to Theorems 3.1 and 3.2 that it provides an upper bound on this complexity). We can apply the following algorithm.

1. For all  $K > 0$ , we choose the model  $\hat{V}(K)$  selected by the penalty  $\text{pen}(V) = K\Delta_V$ .
2. We find  $K_{\min}$  such that  $\Delta_{\hat{V}(K)}$  is very large for  $K < K_{\min}$  and much smaller for  $K > K_{\min}$ .
3. We select  $\hat{V} = \hat{V}(2K_{\min})$ .

The idea is that  $K_{\min}\Delta_V$  shall be the minimal penalty  $\text{pen}_{\min}(V)$  because we observe a jump of the complexity  $\Delta_{\hat{V}}$  around  $K_{\min}\Delta_V$  as expected by **SH1**, **SH2**. Therefore,  $\hat{V}$ , chosen by  $2K_{\min}\Delta_V = 2\text{pen}_{\min}(V)$  shall be optimal from **SH3**.

There exists now several proofs of this heuristic in various problems, see for example [AM09] or [AB10] for the problem of regression on histograms, [Ler11b] and [Ler11a] in density estimation, or [Ver09] for some partial justification of this heuristic in a Gaussian graphical model Selection problem. This section is devoted to the theorems justifying this heuristic in our problem.

#### 4.1 The quadratic loss

**Theorem 4.1.** *Let  $(S, A, P)$  be a random field. Let  $r > 0$ ,  $\epsilon > 0$  and assume that*

$$P \left( \forall V \in \mathcal{V}_s, 0 \leq \text{pen}(V) \leq (1-r) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \right) \geq 1 - \epsilon.$$

Let

$$\hat{V} = \arg \min_{V \in \mathcal{V}_s} \left\{ - \left\| \hat{P}_{i|V} \right\|_{\hat{P}}^2 + \text{pen}(V) \right\},$$

For all  $\delta > 1$ , with probability larger than  $1 - \epsilon - 2\delta^{-1}$ ,

$$\left\| P_{i|\hat{V}} - \hat{P}_{i|\hat{V}} \right\|_{\hat{P}}^2 \geq \sup_{V \in \mathcal{V}_s} \left\{ r \left\| P_{i|V} - \hat{P}_{i|V} \right\|_{\hat{P}}^2 - 2 \left\| P_{i|S} - P_{i|V} \right\|_P^2 \right\} - \frac{17 \ln(N_s^2 \delta)}{3n}.$$

**Comments:**

- When  $V$  is large, the deterministic term  $\left\| P_{i|S} - P_{i|V} \right\|_P^2$  is very small compared to the variance term  $\left\| P_{i|V} - \hat{P}_{i|V} \right\|_{\hat{P}}^2$ . Theorem 4.1 is therefore a minimal penalty theorem. It states that, if the penalty term is too small, the complexity of the selected model (measured here with  $\left\| P_{i|V} - \hat{P}_{i|V} \right\|_{\hat{P}}^2$ ) is as large as possible. This is **SH1** with  $\Delta_V = \left\| P_{i|V} - \hat{P}_{i|V} \right\|_{\hat{P}}^2$ .

Let us now state the associated optimal penalty theorem which proves the slope heuristic.

**Theorem 4.2.** *Let  $(S, A, P)$  be a random field. Let  $\delta > 1$ ,  $r_1 > 0$ ,  $r_2 > 0$ ,  $\epsilon > 0$  and assume that*

$$P \left( \forall V \in \mathcal{V}_s, (1+r_1) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \leq \text{pen}(V) \leq (1+r_2) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \right) \geq 1 - \epsilon.$$

Let

$$\hat{V} = \arg \min_{V \in \mathcal{V}_s} \left\{ - \left\| \hat{P}_{i|V} \right\|_{\hat{P}}^2 + \text{pen}(V) \right\}.$$

For all  $V$  in  $\mathcal{V}_s$ , let  $p_-^V = \inf_{x \in \mathcal{X}(V), P(x(V)) \neq 0} P(x(V))$  and assume that, for some  $\epsilon \leq 1$ ,

$$\inf_{V \in \mathcal{V}_s} p_-^V \geq \epsilon^{-2} \frac{\ln(n N_s \delta)}{n}.$$

Then, there exists an absolute constant  $C$  such that, with probability larger than  $1 - 5\delta^{-1} - \epsilon$ , for all  $V$  in  $\mathcal{V}_s$ , for all  $\eta > 0$ ,

$$\frac{(1-\eta) \wedge (r_1 - C(1+r_1)\epsilon)}{(1+\eta) \vee (r_2 + C(1+r_2)\epsilon)} \left\| P_{i|S} - \hat{P}_{i|\hat{V}} \right\|_P^2 \leq \left\| P_{i|S} - \hat{P}_{i|V} \right\|_P^2 + \frac{6 \ln(N_s^2 \delta)}{\eta n}. \quad (11)$$

**Comments:**

- Let us assume that  $\epsilon \rightarrow 0$ . First, take  $r_1, r_2$  close to 0. The penalty is therefore slightly larger than the minimal penalty  $\left\| P_{i|\hat{V}} - \hat{P}_{i|\hat{V}} \right\|_{\hat{P}}^2$ . It comes from (11) that

$$\left\| P_{i|\hat{V}} - \hat{P}_{i|\hat{V}} \right\|_P^2 \leq C_{r_1, r_2, \eta} \left( \inf_{V \in \mathcal{V}_s} \left\{ \left\| P_{i|S} - \hat{P}_{i|V} \right\|_P^2 \right\} + \frac{6 \ln(N_s^2 \delta)}{\eta n} \right).$$

The complexity of the selected model is therefore the one of an oracle, which is much smaller than the maximal one. We observe a jump of the complexity of the selected model around  $\text{pen}_{\min}$ , this is **SH2**.



- Take  $r_1, r_2$  equal to 1. The penalty is then equal to  $2\text{pen}_{\min}(V)$ . Inequality (11) states, in that case, taking  $\eta$  close to 0, that  $\hat{P}_{i|\hat{V}}$  satisfies an oracle inequality asymptotically optimal. This is **SH3**. We have therefore justified the slope heuristic for the  $L_2$ -risk. In the following section, we give the theorems justifying it for the K ullback loss.

## 4.2 Slope heuristic for the K ullback Loss

The purpose of this section is to give the equivalent of Theorems 4.1 and 4.2 in the case of K ullback loss.

**Theorem 4.3.** *Let  $s > 0$ ,  $\delta > 1$ ,  $\epsilon > 0$ ,  $r > 0$ ,  $p_* > 0$ ,  $\Lambda \geq 100$  and let  $\mathcal{V}_{s,\Lambda,p_*}$ ,  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$  be the collections defined in (3) and (4). For all  $V$  in  $\mathcal{V}_s$ , let*

$$p_2(V) = \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right).$$

Let  $\hat{V}$  be the penalized estimator defined in (9) with a penalty term satisfying

$$P(\forall V \in \mathcal{V}_{s,\Lambda,p_*}, 0 \leq \text{pen}(V) \leq (1-r)p_2(V)) \geq 1 - \epsilon.$$

Then, we have, with probability larger than  $1 - 2\delta^{-1} - \epsilon$ ,

$$p_2(\hat{V}) \geq \max_{V \in \mathcal{V}_{s,\Lambda,p_*}} \{rp_2(V) - 2K(P_{i|S}, P_{i|V})\} - \frac{\ln(N_s^2\delta)}{n} \left( 4 \ln n + \frac{3}{2p_*} \right)$$

Let  $\hat{V}_{(2)}$  be the penalized estimator defined in (10) with a penalty term satisfying

$$P\left(\forall V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}, 0 \leq \text{pen}(V) \leq (1-r)p_2(V)\right) \geq 1 - \epsilon.$$

Also, we have, with probability larger than  $1 - 2\delta^{-1} - \epsilon$ ,

$$p_2(\hat{V}_{(2)}) \geq \max_{V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}} \{rp_2(V) - 2K(P_{i|S}, P_{i|V})\} - \frac{\ln(N_s^2\delta)}{n} \left( 4 \ln n + \frac{3}{2p_*} \right)$$

### Comments:

- Theorem 4.3 states that, when the penalty term is smaller than  $p_2(V)$ , the complexity  $p_2(\hat{V})$  is as large as possible. This is exactly **SH1**, with  $\text{pen}_{\min}(V) = \Delta_V = p_2(V)$ .

**Theorem 4.4.** *Let  $s > 0$ ,  $\delta > 1$ ,  $\epsilon > 0$ ,  $r_1 > 0$ ,  $r_2 > 0$ ,  $p_* > 0$ ,  $\Lambda \geq 100$  and let  $\mathcal{V}_{s,\Lambda,p_*}$ ,  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$  be the collections defined in (3) and (4). For all  $V$  in  $\mathcal{V}_s$ , let  $p_2(V)$  be the quantity defined in Theorem 4.3. Let  $\hat{V}$  be the penalized estimator defined in (9) with a penalty term satisfying*

$$P(\forall V \in \mathcal{V}_{s,\Lambda,p_*}, (1+r_1)p_2(V) \leq \text{pen}(V) \leq (1+r_2)p_2(V)) \geq 1 - \epsilon.$$

Then, there exists an absolute constant  $C$  such that, for all  $\eta > 0$ , with probability larger than  $1 - 2\delta^{-1} - \epsilon$ ,

$$C_L K(P_{i|S}, \hat{P}_{i|\hat{V}}) \leq \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}} \left\{ K(P_{i|S}, \hat{P}_{i|V}) \right\} + \left( 2 \ln(n) + \frac{C_{r_1,r_2,p_*}}{\eta} \right) \frac{\ln(N_s^2\delta)}{n}, \quad (12)$$

where

$$C_L = \frac{(1 - \eta) \wedge (r_1 - C(1 + r_1)\Lambda^{-1/2})}{(1 + \eta) \vee (r_2 + C(1 + r_2)\Lambda^{-1/2})}.$$

Let  $\widehat{V}_{(2)}$  be the penalized estimator defined in (10) with a penalty term satisfying

$$P\left(\forall V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}, (1 + r_1)p_2(V) \leq \text{pen}(V) \leq (1 + r_2)p_2(V)\right) \geq 1 - \epsilon.$$

Also, there exists an absolute constant  $C$  such that, for all  $\eta > 0$ , with probability larger than  $1 - 2\delta^{-1} - \epsilon$ ,

$$C_L K(P_{i|S}, \widehat{P}_{i|\widehat{V}_{(2)}}) \leq \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}^{(2)}} \left\{ K(P_{i|S}, \widehat{P}_{i|V}) \right\} + \left( 2\ln(n) + \frac{C_{r_1,r_2,p_*}}{\eta} \right) \frac{\ln(N_s^2 \delta)}{n}, \quad (13)$$

**Comments:**

- Let us take  $\Lambda = 100 \vee \ln(n)$ . Take at first  $r_1$  and  $r_2$  slightly larger than 0 and therefore a penalty slightly larger than  $\text{pen}_{\min}$ . Then (12) implies that, when  $n$  is sufficiently large  $C_L > 0$ , hence

$$\begin{aligned} p_2(\widehat{V}) &\leq K(P_{i|S}, \widehat{P}_{i|\widehat{V}}) \leq C_L^{-1} \left( \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}} K(P_{i|S}, \widehat{P}_{i|V}) + \left( 2\ln(n) + \frac{C_{r_1,r_2,p_*}}{\eta} \right) \frac{\ln(N_s^2 \delta)}{n} \right) \\ &<< \sup_{V \in \mathcal{V}_{s,\Lambda,p_*}} K(P_{i|V}, \widehat{P}_{i|V}). \end{aligned}$$

This justifies **SH2**.

- Take now  $r_1$  and  $r_2$  equal to 1, so that the penalty is equal to  $2\text{pen}_{\min}$ . Then, we can take  $C_L \rightarrow 1$  in (12). This justifies **SH3**.

## 5 Simulations

In this section we illustrate results obtained in previous sections using simulation experiments. All these simulation experiments can be reproduced by a set of MATLAB<sup>®</sup> routines that can be downloaded from [www.princeton.edu/~dtakahas/publications/LT11routines.zip](http://www.princeton.edu/~dtakahas/publications/LT11routines.zip).

Let  $S = \{-1, 0, 1\} \times \{-1, 0, 1\}$  and  $A = \{-1, 1\}$ . For all the simulations we consider an Ising model on  $A^S$ , with one-point conditional probability for all  $x \in A^S$  given by

$$P_{i|S}(x) = \frac{1}{1 + \exp(-2 \sum_{j \in S} J_{ij} x(i) x(j))}$$

where the pairwise potential  $(J_{ij})_{i,j \in S}$  is given by  $J_{ij} = J \mathbf{1}_{j \in V_i}$  for  $J = 0.2$  and  $V_i \subset G$ . The pair of sites  $(i, j)$  where  $j \in V_i$  is shown in Figure 1. For all these experiments,  $i = (0, 0)$ . We simulated independent samples of the Ising model with increasing sample sizes  $n = 100k$ ,  $k = 1, \dots, 100$ . For each sample size we have  $N = 100$  independent replicas.

### 5.1 Variance term of the risk

The following experiment illustrates Theorem 3.1 and Theorem 3.3. For each sample size we computed the normalized variance terms, namely  $n \left\| \widehat{P}_{i|V_i} - P_{i|V_i} \right\|_P^2$  for the  $L_2$ -norm and  $nK(P_{i|V}, \widehat{P}_{i|V})$  for the K ullback loss. The average values are described in Figure 2 and show that the variance term scales as  $1/n$ . As the behavior of  $L_2$ -norm and K ullback loss is quite similar, in what follows we will show the simulations results only for the  $L_2$ -norm.

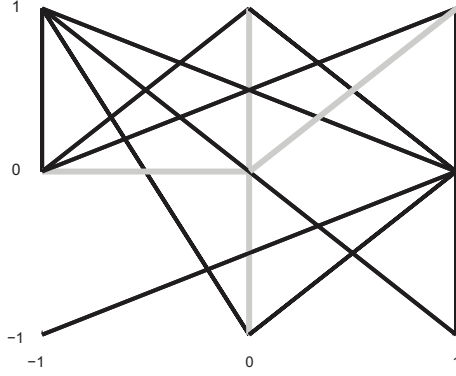


Figure 1: Representation of the interacting pairs of the Ising model used in the simulation experiments. The edges between sites indicate the interacting pairs. The grey colored edges indicate the sites interacting with site  $(0,0)$ .

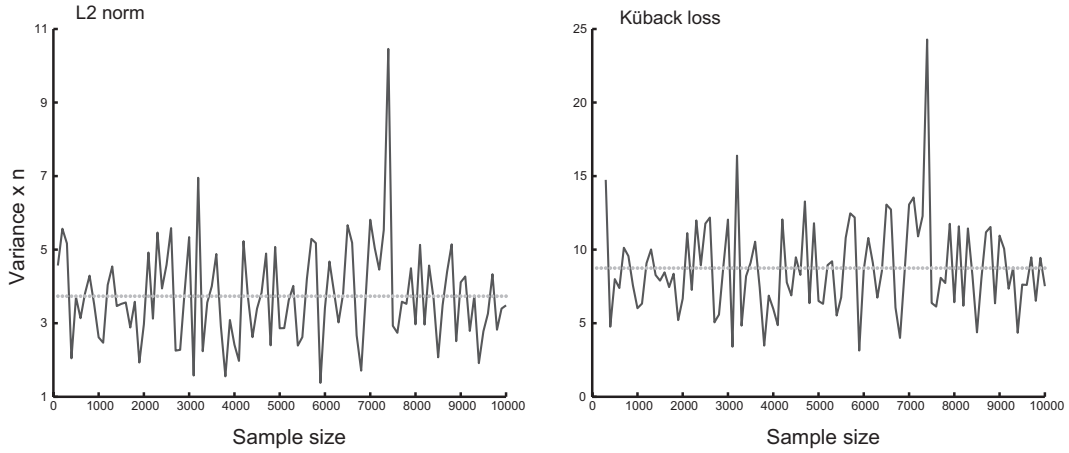


Figure 2: Plot of the number of samples  $n$  against  $n \left\| \hat{P}_{i|V_i} - P_{i|V_i} \right\|_P^2$  for the  $L_2$  norm and  $nK(P_{i|V}, \hat{P}_{i|V})$  for the Küback loss. The dotted lines indicate the linear regression lines. Observe that the regression line is essentially parallel to the abscissa.

## 5.2 Slope heuristic

Here we illustrate the slope heuristic. We use 500 samples from the Ising model described in the beginning of this section. We use as the measure of complexity, for  $i = (0,0)$  and  $V \subset S$ , the quantity  $\|P_{i|V} - \hat{P}_{i|V}\|_{\hat{P}}^2$ . In Figure 3 we plot the value of the measure of complexity against the criterion

$$\min_{V \in S} \{-\|\hat{P}_{i|V}\|_{\hat{P}}^2 + c\|P_{i|V} - \hat{P}_{i|V}\|_{\hat{P}}^2\},$$

for the positive constants  $c < 8$ . We clearly see that when  $c$  is smaller than 1 the complexity is the largest possible and this is the content of Theorem 4.1. We also observe that when  $c$  is slightly larger than 1 there is a sudden decrease in the complexity, which is the content of Theorem 4.2. Finally, the model chosen by  $c = 2$  is exactly the one given by oracle as predicted by Theorem 4.2.

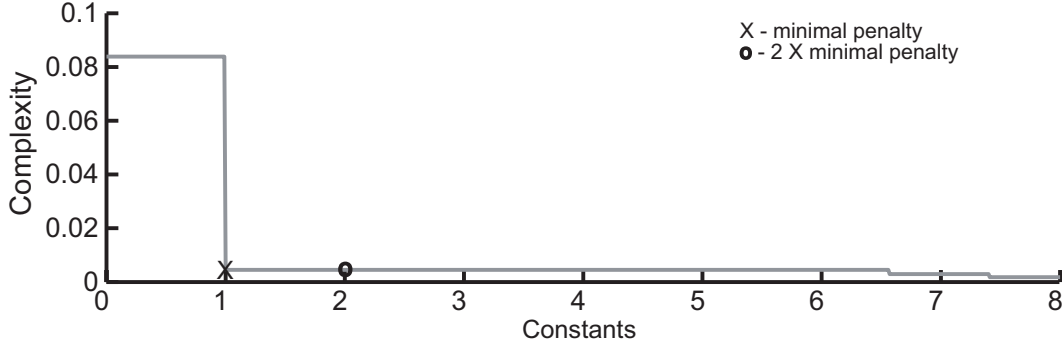


Figure 3: Example of slope heuristic. Observe the sudden change in behavior around the minimal penalty.

### 5.3 Oracle risk compared to the risk of the estimated model

Observe that, in the simulation above, we used the quantity  $c\|P_{i|V} - \hat{P}_{i|V}\|_P^2$  as the penalty term. In practice we cannot compute this quantity and we use instead the quantity  $ca^{v-1}n^{-1}$  given in Theorem 3.2. Here we will illustrate the performances of the slope heuristic using this last quantity. Let  $\hat{V}(2\tilde{C}_{\min})$  be the neighborhood selected by the slope heuristic. One way to verify the performances of the slope heuristic is to compute the risk ratio

$$\frac{\left\| \hat{P}_{i|\hat{V}(2\tilde{C}_{\min})} - P_{i|S} \right\|_P^2}{\inf_{V \subset S} \left\| \hat{P}_{i|V} - P_{i|S} \right\|_P^2}. \quad (14)$$

For each sample size, we computed the ratio (14) for 100 different samples and we obtained the average. The result is summarized in Figure 4. For comparison, we estimated also the average risk ratio for the model selected using the theoretical constant  $6Ka^{-1}$  with  $K = 2$  given by Theorem 3.2. Observe that when the sample size  $n$  increases, the risk ratio of the model estimated by the slope heuristic approximates one, as we expect from Theorem 4.2. Also, we observe that the slope heuristic has in general a better risk compared to the criteria using the theoretical constant.

## 6 Discussion

The problem of recovering the support  $V^* \subset S$  of  $P_{i|S}$  is an active area of research (see [BMS08, CT06a, GOT10, RWL10]). The main drawback of these works is the restrictions imposed to guarantee the results. In particular, it is always assumed that all the sites of interest are observed, *i.e.*,  $V_M = S$ . This is never the case in important applications like neuroscience and molecular biology. In neuroscience, for example, the experimenter has only access to a tiny fraction of the whole neural network and has to make inferences based on it. Clearly the exact recovery of  $V^*$  is out of question, but rather a good approximation to the local rules  $P_{i|S}$  is desired. The model selection approach is a natural way to formulate this problem.

We may wonder if the conditions in [BMS08, CT06a, GOT10, RWL10] are satisfied if the measure  $P$  of interest is not the one on  $A^S$  but the projection on  $A^{V_M}$ . Unfortunately, this is also not the case, because [BMS08, CT06a, GOT10, RWL10] assumed that  $P$  is Gibbsian and, in general, a projection of a Gibbs measure is not Gibbsian [FP97].

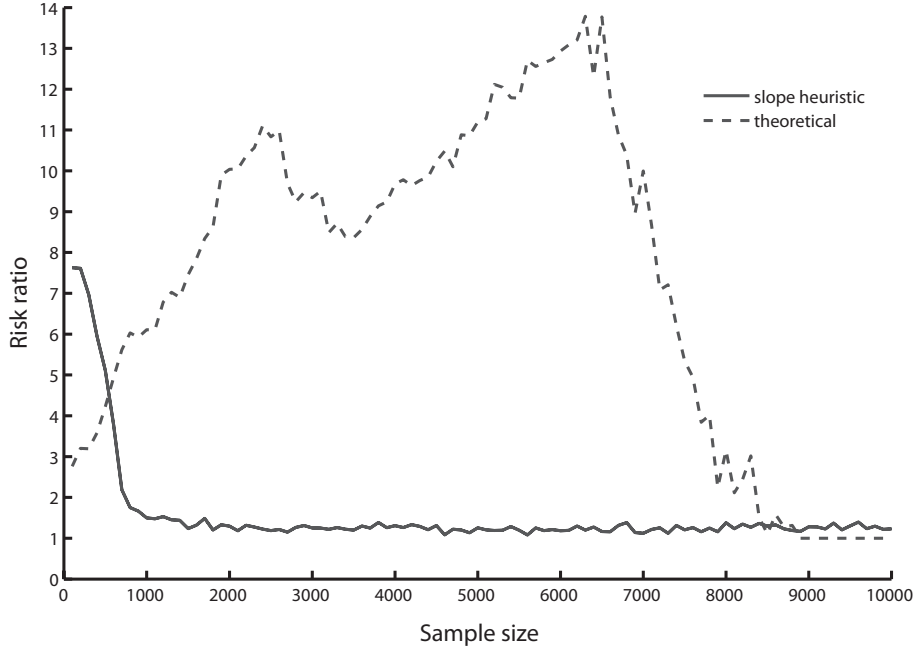


Figure 4: Plot of the number of sample size  $n$  against the average of risk ratio for the models selected by the slope heuristic (solid line) and by the theoretical constant (dashed line).

We think that these considerations alone are important enough to justify our work on model selection procedures for general random fields, but a more detailed comparisons emphasizes our points.

[CT06a] considered the problem of recovering the support of  $P_{i|S}$  in an homogeneous, finite range random field based on one realization. The homogeneity is not realistic in our applications and the comparison with our work is not straightforward, nevertheless we observe some interesting aspects.

1. The consistency result in [CT06a] is asymptotic whereas all our results are non-asymptotic and hold for all  $n$ .
2. They considered finite range interaction random fields eventually included into the observed sites. Our approach let us work with non-observed sites and infinite range random fields.
3. The number of observed sites  $|\Lambda|$  in [CT06a] is the analogous quantity for the number of samples  $n$  in this paper. Theorem 2.1 in their article shows that they select a neighborhood of order  $o(\log^{1/2} n)$  among the  $o(\log^{1/2} n)$  closest sites. Our model selection algorithm can be applied in high dimension situations and allows maximum neighborhood size of  $O(\log n)$  selected from  $O(e^{n^\beta})$ ,  $0 \leq \beta < 1$ , possible sites.
4. They considered penalized log-likelihood estimators as those that we studied in the Küllback case. Our results on Küllback loss can therefore be seen as natural extensions of those in [CT06a] for the model selection setup. Our penalty, designed for the oracle approach, is of AIC-type  $Ka^v/n$  whereas they considered, for exact recovery, a BIC-type penalty of order  $Ka^v \ln n/n$ . This is a difference between the oracle approach and model identification that was already noticed in a regression framework, see for example [Yan05].

[GOT10] considered the problem of recovering the support of  $P_{i|S}$  for infinite range Ising models in  $\mathbb{Z}^d$ . The main restrictions in this work are that the interactions between the sites are supposed to be pairwise, weak (“high temperature”) and that a subset of the observed sites of size  $O(\log(n))$ , where  $n$  is the sample size, must be fixed to apply the proposed procedure. Our procedure has no restriction on the strength of interaction, can be applied for non-pairwise interactions, and we do not need to fix a subset of observed sites.

In [BMS08], the analysis is restricted to finite random fields, where the maximum neighborhood size is known a priori. For infinite range random fields, these results are useless since the “constants”  $\epsilon$  and  $\delta$ , that should be positive, are both equal to 0 in general. More importantly, the procedure used the knowledge of the lower bound  $\epsilon$  on the bias term. As this  $\epsilon$  is unknown in practice, it is not clear, even if the underlying model is a finite random fields, how it should be evaluated. It is not clear how to generalize these results to the case where the maximum neighborhood size is allowed to grow with  $n$ . This would require a careful analysis of the behavior of the quantities  $\epsilon$  and  $\delta$  which are hard to compute even in simple models. Nevertheless, in the specific case when the underlying random field is the Ising model, a straightforward computation Theorem 3 in [BMS08] shows that, when the number of total sites is  $O(e^{n^\beta})$ ,  $0 \leq \beta < 1$ , the maximum size of the allowed neighborhood is  $O(\log n)$ .

In [LT11], we introduced a model selection procedure for  $L_\infty$ -risk. We worked with random fields with binary alphabet and under some restrictions on the probability measure  $P$ . We showed the superiority of the oracle approach compared to the identification procedures available in the literature. However, we were not able to prove the slope heuristic. In the present work, we obtained sharper oracle inequalities, we proved the slope heuristic and removed all the restrictions of [LT11].

The proof of the slope heuristic for general model selection problems is still in its beginning and our results of Section 4 are major contributions to this problem. In particular, we provide, up to our knowledge, the first proofs of this heuristic in a discrete framework. Moreover, our proof in the Küllback case is the only one with [Sau11] that holds for a non-Hilbertian risk. Finally, following the notations of [AM09], the proofs usually rely on good concentration properties of the terms  $p_1$  and  $p_2$  and a comparison of their expectations. We proceed here with a direct comparison of these terms, proving some typicality results for the terms  $p_1$  and  $p_2$ . See Theorem 8.7 and Lemma 8.22. Our approach can be understood as a pathwise version of the strategy suggested in [AM09].

The work [RWL10] is restricted to the Ising model on finite graphs and assumes the incoherence condition, which is a very restrictive (see [BM09]). Nevertheless, the use of  $\ell_1$ -penalization allows a computationally efficient implementation of the algorithm proposed in [RWL10]. This is critical in applications. For the moment, our algorithm lacks the computational efficiency. To have a fast implementation of our algorithm or of an approximation of it will be our main task in a future work.

We provide in Table 1 a comparative summary of the available results.

## Acknowledgements

We are grateful to Antonio Galves for many discussions and fruitful advices during the redaction of the paper.

ML was supported by FAPESP grant 2009/09494-0. DYT was supported by FAPESP grant 2008/08171-0 and by Pew Latin American Fellowship. This work is part of USP project “Mathematics, computation, language and the brain”.

Table 1: Compative table of related results

Algorithm	This work	[BMS08]	[CT06a]	[GOT10]	[RWL10]	[LT11]
Infinite range interactions	yes	no	no	yes	no	yes
Partially observed interacting sites	yes	yes, strong conditions	no	yes, high temperature	no	yes
Non-binary alphabet	yes	yes	yes	no	no	no
Non-pairwise interaction	yes	yes	yes	no	no	no
Ising model on $\mathbb{Z}^d$ below critical temperature	yes	no	yes	no	no	yes
Restrictions on the interaction graph	none	none	none	none	Incoherence	none
Maximal size of the neighborhoods	$O(\ln n)$	$O(\ln n)$	$o(\ln n)$	$O(\ln n)$	$O(n^\alpha)$	$O(\ln n)$
Number of possible sites for the candidate neighborhoods	$O(e^{n^\beta})$	$O(e^{n^\beta})$	$o(\ln n)$	$O(\ln n)$	$O(e^{n^\beta})$	$O(e^{n^\beta})$

## 7 Proofs

### 7.1 Proof of Theorem 3.1:

Let  $\theta > 0$  to be chosen later and let  $Q$  denote either  $P$  or  $\hat{P}$ . We decompose the risk as follows

$$\begin{aligned}
\left\| \hat{P}_{i|V} - P_{i|V} \right\|_Q^2 &= \sum_{x \in \mathcal{X}(V)} \frac{Q(x(V/\{i\}))}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2 \\
&= \sum_{x \in \mathcal{X}(V), Q(x(V/\{i\})) \leq \theta(a^n)^{-1}} \frac{Q(x(V/\{i\}))}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2 \\
&\quad + \sum_{x \in \mathcal{X}(V), Q(x(V/\{i\})) > \theta(a^n)^{-1}} \frac{Q(x(V/\{i\}))}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2
\end{aligned}$$

As the cardinal of  $\mathcal{X}(V)$  is  $a^n$  and  $\left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2 \leq 1$ , the first term in this decomposition is upper bounded by  $\theta n^{-1}$ . Hence

$$\left\| \hat{P}_{i|V} - P_{i|V} \right\|_Q^2 = \frac{\theta}{n} + \sum_{x \in \mathcal{X}(V), Q(x(V/\{i\})) > \theta(a^n)^{-1}} \frac{Q(x(V/\{i\}))}{a} \left( \hat{P}_{i|V} - P_{i|V} \right)^2 \quad (15)$$

Hereafter in the proof of Theorem 3.1, we denote by

$$\mathcal{X}^\theta(V) = \{x \in \mathcal{X}(V), Q(x(V/\{i\})) > \theta(a^v n)^{-1}\}.$$

It comes from Lemma 8.1 that

$$\begin{aligned} \|\hat{P}_{i|V} - P_{i|V}\|_P^2 - \frac{\theta}{n} &= \sum_{x \in \mathcal{X}^\theta(V)} \frac{P(x(V/\{i\}))}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2 \\ &\leq \sum_{x \in \mathcal{X}^\theta(V)} \frac{\left( \left| \hat{P}(x(V)) - P(x(V)) \right| + \hat{P}_{i|V}(x) \left| \left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right) \right| \right)^2}{aP(x(V/\{i\}))} \\ &\leq \frac{2}{a} \left( \sum_{x \in \mathcal{X}^\theta(V)} \frac{\left( \hat{P}(x(V)) - P(x(V)) \right)^2}{P(x(V/\{i\}))} + \sum_{x \in \mathcal{X}^\theta(V/\{i\})} \frac{\left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))} \right). \end{aligned}$$

From Lemma 8.1, we also have

$$|\hat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left| \hat{P}(x(V)) - P(x(V)) \right| + P_{i|V}(x) \left| \left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right) \right|}{a\hat{P}(x(V/\{i\}))}.$$

Hence

$$|\hat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left| \hat{P}(x(V)) - P(x(V)) \right| + (P_{i|V}(x) + \hat{P}_{i|V}(x)) \left| \left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right) \right|}{a\sqrt{\hat{P}(x(V/\{i\}))P(x(V/\{i\}))}}.$$

Thus,

$$\|\hat{P}_{i|V} - P_{i|V}\|_{\hat{P}}^2 - \frac{\theta}{n} = \sum_{x \in \mathcal{X}^\theta(V)} \frac{\hat{P}(x(V/\{i\}))}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2$$

is smaller than

$$\begin{aligned} &\sum_{x \in \mathcal{X}^\theta(V)} \frac{\left( \left| \hat{P}(x(V)) - P(x(V)) \right| + (\hat{P}_{i|V}(x) + P_{i|V}(x)) \left| \left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right) \right| \right)^2}{aP(x(V/\{i\}))} \\ &\leq \frac{2}{a} \left( \sum_{x \in \mathcal{X}^\theta(V)} \frac{\left( \hat{P}(x(V)) - P(x(V)) \right)^2}{P(x(V/\{i\}))} + 2 \sum_{x \in \mathcal{X}^\theta(V/\{i\})} \frac{\left( \hat{P}(x(V/\{i\})) - P(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))} \right). \end{aligned}$$

We use Theorem 8.14 with  $b = \sqrt{\theta^{-1}a^v n}$ , for all  $x > 0$ , for all  $\eta > 0$ , we have, with probability larger than  $1 - 2e^{-x}$ ,

$$\|\hat{P}_{i|V} - P_{i|V}\|_Q^2 \leq \frac{\theta}{n} + \frac{6}{a} \left( (1 + \eta)^3 \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{32a^v x^2}{\theta \eta^3 n} \right).$$

Take  $\theta = 8a^{v/2}x\eta^{-3/2}$ , we obtain

$$\|\hat{P}_{i|V} - P_{i|V}\|_Q^2 \leq \frac{6}{a} \left( (1 + \eta)^3 \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{6a^{v/2}x}{\eta^{3/2}n} \right).$$

Using  $ab \leq \eta a^2 + (4\eta)^{-1}b^2$ , we finally get

$$\|\hat{P}_{i|V} - P_{i|V}\|_Q^2 \leq \frac{6}{a} \left( (1 + 8\eta) \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{9x^2}{\eta^4 n} \right).$$



## 7.2 Proof of Theorem 3.2:

For all probability measures  $Q$ , let  $(\cdot, \cdot)_Q$  be the scalar product associated to the  $L_{2,Q}$ -norm  $\|\cdot\|_Q$ . Let  $V$  and  $V'$  in the collection  $\mathcal{V}_s$ . We have

$$\begin{aligned} \frac{1}{a} \sum_{x \in \mathcal{X}(V \cup V')} \hat{P}(x(V \cup V')) P_{i|V}(x) \\ = \sum_{x \in \mathcal{X}(V)} \frac{\hat{P}(x(V/\{i\}))}{a} \hat{P}_{i|V}(x) P_{i|V}(x) = \left( \hat{P}_{i|V}, P_{i|V} \right)_{\hat{P}}. \\ \frac{1}{a} \sum_{x \in \mathcal{X}(V \cup V')} P(x(V \cup V')) P_{i|V}(x) = \sum_{x \in \mathcal{X}(V)} \frac{P(x(V/\{i\}))}{a} P_{i|V}^2(x) = \|P_{i|V}\|_P^2 \end{aligned}$$

Hence, for all  $V, V'$  in  $\mathcal{V}_s$ ,

$$\begin{aligned} \|\hat{P}_{i|V}\|_{\hat{P}}^2 &= \|P_{i|V}\|_{\hat{P}}^2 + 2 \left( \hat{P}_{i|V} - P_{i|V}, P_{i|V} \right)_{\hat{P}} + \|\hat{P}_{i|V} - P_{i|V}\|_{\hat{P}}^2 \\ &= \|P_{i|V}\|_P^2 + \|\hat{P}_{i|V} - P_{i|V}\|_{\hat{P}}^2 - \left( \|P_{i|V}\|_{\hat{P}}^2 - \|P_{i|V}\|_P^2 \right) \\ &\quad + \frac{2}{a} \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) P_{i|V}(x). \end{aligned} \tag{16}$$

Moreover, from Pythagoras relation see Proposition 8.17, we have

$$\|P_{i|S} - P_{i|V}\|_P^2 = \|P_{i|S}\|_P^2 - \|P_{i|V}\|_P^2.$$

By definition of  $\hat{V}$ , we have, for all  $V$  in  $\mathcal{V}_s$ ,

$$\|P_{i|S}\|_P^2 - \|\hat{P}_{i|\hat{V}}\|_{\hat{P}}^2 + \text{pen}(\hat{V}) \leq \|P_{i|S}\|_P^2 - \|\hat{P}_{i|V}\|_{\hat{P}}^2 + \text{pen}(V)$$

Hence, for all  $0 < \nu \leq 1$ , from (16),

$$\nu \|\hat{P}_{i|V} - P_{i|V}\|_{\hat{P}}^2 \leq \|P_{i|S} - P_{i|\hat{V}}\|_P^2 + \nu \|\hat{P}_{i|\hat{V}} - P_{i|\hat{V}}\|_P^2$$

is smaller than

$$\begin{aligned} &\|P_{i|S} - P_{i|V}\|_P^2 + \text{pen}(V) - \|\hat{P}_{i|V} - P_{i|V}\|_{\hat{P}}^2 - \left( \text{pen}(\hat{V}) - \|\hat{P}_{i|\hat{V}} - P_{i|\hat{V}}\|_{\hat{P}}^2 - \nu \|\hat{P}_{i|\hat{V}} - P_{i|\hat{V}}\|_P^2 \right) \\ &+ \left( \|P_{i|V}\|_{\hat{P}}^2 - \|P_{i|V}\|_P^2 - \|P_{i|\hat{V}}\|_{\hat{P}}^2 + \|P_{i|\hat{V}}\|_P^2 \right) \\ &+ \frac{2}{a} \sum_{x \in \mathcal{X}(V \cup \hat{V})} (\hat{P}(x(V \cup \hat{V})) - P(x(V \cup \hat{V}))) \left( P_{i|\hat{V}}(x) - P_{i|V}(x) \right). \end{aligned} \tag{17}$$

We have also,

$$\begin{aligned} &\|P_{i|V}\|_{\hat{P}}^2 - \|P_{i|V}\|_P^2 - \|P_{i|\hat{V}}\|_{\hat{P}}^2 + \|P_{i|\hat{V}}\|_P^2 \\ &= \frac{1}{a} \sum_{x \in \mathcal{X}((V \cup \hat{V})/\{i\})} (\hat{P}(x((V \cup \hat{V})/\{i\})) - P(x((V \cup \hat{V})/\{i\}))) \left( P_{i|V}^2(x) - P_{i|\hat{V}}^2(x) \right). \end{aligned}$$

Let  $0 < \eta \leq 1$ ,  $\delta > 1$  and assume that,  $N_s \geq 2$ . Let  $\Omega^\delta$  be the intersection of the following events:

$$\Omega_1^\delta = \left\{ \forall V \in \mathcal{V}_s, \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \leq \frac{6}{a} \left( (1 + 8\eta) \frac{a^v}{n} + \frac{13 \ln(2N_s \delta)^2}{\eta^4 n} \right) \right\}.$$

$$\Omega_2^\delta = \left\{ \forall V \in \mathcal{V}_s, \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \leq \frac{6}{a} \left( (1 + 8\eta) \frac{a^v}{n} + \frac{13 \ln(2N_s \delta)^2}{\eta^4 n} \right) \right\}.$$

$$\begin{aligned} \Omega_3^\delta = \left\{ \forall V, V' \in \mathcal{V}_s^2, \left\| P_{i|V} \right\|_{\hat{P}}^2 - \left\| P_{i|V} \right\|_P^2 - \left\| P_{i|V'} \right\|_{\hat{P}}^2 + \left\| P_{i|V'} \right\|_P^2 \right. \\ \left. \leq 2 \left\| P_{i|V} - P_{i|V'} \right\|_P \sqrt{2 \frac{\ln(N_s^2 \delta)}{n}} + \frac{\ln(N_s^2 \delta)}{3n} \right\}. \end{aligned} \quad (18)$$

$$\begin{aligned} \Omega_4^\delta = \left\{ \forall V, V' \in \mathcal{V}_s^2, \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \frac{P_{i|V'}(x) - P_{i|V}(x)}{a} \right. \\ \left. \leq \left\| P_{i|V} - P_{i|V'} \right\|_P \sqrt{2 \frac{\ln(N_s^2 \delta)}{n}} + \frac{\ln(N_s^2 \delta)}{3n} \right\}. \end{aligned} \quad (19)$$

Theorem 3.1, Lemma 8.16 and union bounds give that

$$P \left( \left( \Omega^\delta \right)^c \right) \leq \frac{4}{\delta}.$$

For all  $V, V'$  in  $\mathcal{V}_s$  and all  $\xi > 0$ , on  $\Omega^\delta$ , we have

$$\begin{aligned} 2 \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \frac{P_{i|V'}(x) - P_{i|V}(x)}{a} + \left\| P_{i|V} \right\|_{\hat{P}}^2 \\ - \left\| P_{i|V} \right\|_P^2 - \left\| P_{i|V'} \right\|_{\hat{P}}^2 + \left\| P_{i|V'} \right\|_P^2 \leq \frac{\xi}{2} \left\| P_{i|V} - P_{i|V'} \right\|_P^2 + \left( \frac{16}{\xi} + 1 \right) \frac{\ln(N_s^2 \delta)}{3n}. \end{aligned}$$

From (17), we deduce that, on  $\Omega^\delta$ , for all  $0 < \xi < \eta$ ,

$$\begin{aligned} (\nu - \xi) \left\| P_{i|S} - \hat{P}_{i|\hat{V}} \right\|_P^2 &\leq (1 + \xi) \left\| P_{i|S} - P_{i|V} \right\|_P^2 + \text{pen}(V) \\ &\quad - \left( \text{pen}(\hat{V}) - (1 + \nu)(1 + \eta)^3 \frac{6}{a} \frac{a^{\hat{v}}}{n} \right) \\ &\quad + \frac{1}{n} \left( \frac{78(1 + \nu)}{\eta^4 a} (\ln(2N_s \delta))^2 + \left( \frac{16}{\xi} + 1 \right) \ln(N_s^2 \delta) \right). \end{aligned}$$

Take at first  $0 < \xi < \nu$  and  $0 < \eta$  sufficiently small to ensure that  $(1 + \nu)(1 + \eta)^3 \leq K$  to obtain (7). To obtain (8), choose  $\nu = 1$  and  $\eta > 0$  sufficiently small to ensure that  $(1 + \eta)^3 < K/2$  and  $\xi = (\ln(N_s^2 \delta))^{-1}$ . We conclude the proof, saying that the inequality is obvious when  $\delta < 4$ , and, when  $\delta \geq 4$ ,

$$\frac{1 + (\ln N_s^2 \delta)^{-1}}{1 - (\ln N_s^2 \delta)^{-1}} = 1 + \frac{2(\ln N_s^2 \delta)^{-1}}{1 - (\ln N_s^2 \delta)^{-1}} \leq 1 + \frac{2(\ln \delta)^{-1}}{1 - (\ln \delta)^{-1}} \leq 1 + \frac{8}{\ln \delta}.$$

### 7.3 Proof of Theorem 3.3:

Let  $V$  in  $\mathcal{V}_{s,\Lambda}$  or  $\mathcal{V}_{s,\Lambda}^{(2)}$  and let us define

$$p_1(V) = \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\hat{P}_{i|V}(x)} \right), \quad p_2(V) = \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right).$$

From Lemma 8.21 and we have

$$\begin{aligned} p_1(V) &\leq \frac{10}{3} \sum_{x \in \mathcal{X}(V)} \frac{\left( P(x(V)) - \hat{P}(x(V)) \right)^2}{P(x(V))} + \frac{14}{9} \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left( P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))} \\ p_2(V) &\leq \frac{3}{2} \sum_{x \in \mathcal{X}(V)} \frac{\left( P(x(V)) - \hat{P}(x(V)) \right)^2}{P(x(V))} + \frac{7}{3} \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left( P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))}. \end{aligned}$$

Let  $V_* = V$  or  $V/\{i\}$ . On the event  $\Omega_{prob}(\delta)$  defined in Lemma 8.20, thanks to Lemma 8.20, we have

$$\sup_{x \in \mathcal{X}(V_*)} \frac{1}{\sqrt{P(x(V_*))}} \leq \sup_{x \in \mathcal{X}(V_*)} \frac{1}{\sqrt{P(x(V_*))}} \leq \sup_{x \in \mathcal{X}(V_*)} \frac{1 + 2\Lambda^{-1/2}}{\sqrt{\hat{P}(x(V_*))}} \leq \frac{2\sqrt{n}}{\sqrt{\Lambda \ln(2a^s N_s \delta)}}.$$

As this quantity is not random, the same bound holds on  $\Omega_{prob}(\delta)^c$ . We can apply Theorem 8.14 to get that, for all  $x > 0$ , for all  $\eta > 0$ , with probability larger than  $1 - 2e^{-x}$ ,

$$\begin{aligned} p_1(V) &\leq \frac{44}{9} \left( (1 + \eta)^3 \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{128x^2}{n\eta^3 \Lambda \ln(2a^s N_s \delta)} \right). \\ p_2(V) &\leq \frac{23}{6} \left( (1 + \eta)^3 \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{128x^2}{n\eta^3 \Lambda \ln(2a^s N_s \delta)} \right). \end{aligned}$$

We use a union bound to obtain that, for all  $V$  in  $\mathcal{V}_{s,\Lambda}$  or  $\mathcal{V}_{s,\Lambda}^{(2)}$ , with probability larger than  $1 - \delta$ ,

$$\begin{aligned} p_1(V) &\leq \frac{44}{9} \left( (1 + \eta)^3 \frac{a^v}{n} + \left( \frac{4}{\eta} + \frac{128}{\eta^3 \Lambda} \right) \frac{\ln(2N_s \delta)}{n} \right). \\ p_2(V) &\leq \frac{23}{6} \left( (1 + \eta)^3 \frac{a^v}{n} + \left( \frac{4}{\eta} + \frac{128}{\eta^3 \Lambda} \right) \frac{\ln(2N_s \delta)}{n} \right). \end{aligned}$$

### 7.4 Proof of Theorem 3.4:

Let us first decompose the selection criterion as follows.

$$\begin{aligned} - \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \hat{P}_{i|V}(x) \right) + \text{pen}(V) &= K(P_{i|S}, \hat{P}_{i|V}) + \text{pen}(V) - p_1(V) - p_2(V) + L(V) \\ &\quad + \int dP(x(S)) \ln \left( \frac{1}{P_{i|S}(x)} \right). \end{aligned} \tag{20}$$

In the previous decomposition, we have

$$\begin{aligned} p_1(V) &= \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\hat{P}_{i|V}(x)} \right). \\ p_2(V) &= \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right). \\ L(V) &= \sum_{x \in \mathcal{X}(V)} (\hat{P}(x(V)) - P(x(V))) \ln \left( \frac{1}{P_{i|V}(x)} \right). \end{aligned}$$

We deduce from (20) and the definition of  $\hat{V}$  that, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$K(P_{i|S}, \hat{P}_{i|\hat{V}}) \leq K(P_{i|S}, P_{i|V}) + \text{pen}(V) - p_2(V) - \left( \text{pen}(\hat{V}) - p_1(\hat{V}) - p_2(\hat{V}) \right) + L(V) - L(\hat{V}). \quad (21)$$

Let  $\Omega_{\text{prob}}(\delta)$  defined in Lemma 8.20. Let  $\eta > 0$  and let  $\Omega_{p1,p2}(\delta)$  be the event, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$

$$\begin{aligned} p_1(V) &\leq 5 \left( (1+\eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \\ p_2(V) &\leq 4 \left( (1+\eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\epsilon^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \end{aligned}$$

Let  $\Omega_L(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$(L(V) - L(V')) \mathbf{1}_{\Omega_{\text{prob}}(\delta)} \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).$$

Let  $\Omega = \Omega_{\text{prob}}(\delta) \cap \Omega_{p1,p2}(\delta) \cap \Omega_L(\delta)$ . It comes from Lemma 8.20, Theorem 3.3 and Lemma 8.23 that  $P(\Omega^c) \leq 3\delta$ . Moreover, on  $\Omega$ , we have, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$\begin{aligned} p_1(\hat{V}) + p_2(\hat{V}) + L(V) - L(\hat{V}) &\leq \text{pen}(\hat{V}) + \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|\hat{V}})) \\ &\quad + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3} \Lambda} \right). \end{aligned}$$

Hence, on  $\Omega$ ,

$$\frac{1-\eta}{1+\eta} K(P_{i|S}, \hat{P}_{i|\hat{V}}) \leq K(P_{i|S}, P_{i|V}) + \text{pen}(V) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3} \Lambda} \right).$$

We deduce from (20) and the definition of  $\hat{V}_{(2)}$  that, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$K(P_{i|S}, \hat{P}_{i|\hat{V}_{(2)}}) \leq K(P_{i|S}, P_{i|V}) + \text{pen}(V) - p_2(V) - \left( \text{pen}(\hat{V}_{(2)}) - p_1(\hat{V}_{(2)}) - p_2(\hat{V}_{(2)}) \right) + L(V) - L(\hat{V}_{(2)}). \quad (22)$$

Let  $\Omega_{\text{prob}}(\delta)$  defined in Lemma 8.20. Let  $\eta > 0$  and let  $\Omega_{p1,p2}^{(2)}(\delta)$  be the event, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$

$$\begin{aligned} p_1(V) &\leq 5 \left( (1+\eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\eta^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \\ p_2(V) &\leq 4 \left( (1+\eta)^3 \frac{a^v}{n} + \left( 1 + \frac{64}{\epsilon^2 \Lambda} \right) \frac{4 \ln(2N_s \delta)}{\eta n} \right). \end{aligned}$$

Let  $\Omega_L^{(2)}(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$(L(V) - L(V'))\mathbf{1}_{\Omega_{prob}(\delta)} \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2\delta)}{n} \left( 4\ln n + \frac{3}{2\eta p_*} \right).$$

Let  $\Omega = \Omega_{prob}(\delta) \cap \Omega_{p1,p2}^{(2)}(\delta) \cap \Omega_L^{(2)}(\delta)$ . It comes from Lemma 8.20, Theorem 3.3 and Lemma 8.23 that  $P(\Omega^{(2)}) \geq 1 - 3\delta$ . Moreover, on  $\Omega^{(2)}$ , we have, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$\begin{aligned} p_1(\widehat{V}_{(2)}) + p_2(\widehat{V}_{(2)}) + L(V) - L(\widehat{V}_{(2)}) &\leq \text{pen}(\widehat{V}_{(2)}) + \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|\widehat{V}_{(2)}})) \\ &\quad + \frac{\ln(N_s^2\delta)}{n} \left( 4\ln n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3}\Lambda} \right). \end{aligned}$$

Hence, on  $\Omega^{(2)}$ ,

$$\frac{1-\eta}{1+\eta} K(P_{i|S}, \widehat{P}_{i|\widehat{V}_{(2)}}) \leq K(P_{i|S}, P_{i|V}) + \text{pen}(V) + \frac{\ln(N_s^2\delta)}{n} \left( 4\ln n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3}\Lambda} \right).$$

## 7.5 Proof of Theorem 4.1:

Let us introduce, for all  $V$  in  $\mathcal{V}_s$ ,

$$L(V) = \|P_{i|V}\|_P^2 - \|P_{i|V}\|_{\widehat{P}}^2 + \frac{2}{a} \sum_{x \in \mathcal{X}(V)} (\widehat{P}(x(V)) - P(x(V)))P_{i|V}(x).$$

By definition of  $\widehat{V}$ , we have, for all  $V$  in  $\mathcal{V}_s$ ,

$$\|P_{i|S}\|_P^2 - \|\widehat{P}_{i|\widehat{V}}\|_{\widehat{P}}^2 + \text{pen}(\widehat{V}) \leq \|P_{i|S}\|_P^2 - \|\widehat{P}_{i|V}\|_{\widehat{P}}^2 + \text{pen}(V).$$

Hence from inequality (16) in the proof of Theorem 3.2, we have, for all  $V$  in  $\mathcal{V}_s$ ,

$$\begin{aligned} &\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \left( \text{pen}(\widehat{V}) - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2 \right) - L(\widehat{V}) \\ &\leq \|P_{i|S} - P_{i|V}\|_P^2 + \left( \text{pen}(V) - \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \right) - L(V). \end{aligned} \quad (23)$$

Let  $\Omega_{\text{pen}} = \left\{ 0 \leq \text{pen}(V) \leq (1-r) \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \right\}$  and let  $\Omega_{\min \text{pen}}^\delta = \Omega_3^\delta \cap \Omega_4^\delta \cap \Omega_{\text{pen}}$ , where  $\Omega_3^\delta$  and  $\Omega_4^\delta$  are respectively defined in (18) and (19). It comes from Lemma 8.16 and our assumption on  $\text{pen}(V)$  that  $P((\Omega_{\min \text{pen}}^\delta)^c) \leq \epsilon + 2\delta^{-1}$ . Moreover, on  $\Omega_{\min \text{pen}}^\delta$ , we have, for all  $\eta > 0$ ,

$$|L(\widehat{V}) - L(V)| \leq \eta \|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \eta \|P_{i|S} - P_{i|V}\|_P^2 + \left( \frac{16}{\eta} + 1 \right) \frac{\ln(N_s^2\delta)}{3n}.$$

$$\begin{aligned} &(1-\eta) \|P_{i|S} - P_{i|\widehat{V}}\|_P^2 - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2 \\ &\leq (1+\eta) \|P_{i|S} - P_{i|V}\|_P^2 - r \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 + \left( \frac{16}{\eta} + 1 \right) \frac{\ln(N_s^2\delta)}{3n}. \end{aligned}$$

We conclude the proof choosing  $\eta = 1$ .

## 7.6 Proof of Theorem 4.2:

Let

$$\Omega_{\text{pen}} = \left\{ \forall V \in \mathcal{V}_s, (1+r_1) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \leq \text{pen}(V) \leq (1+r_2) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \right\},$$

let  $\Omega_{\text{comp}}^\delta = \Omega_3^\delta \cap \Omega_4^\delta \cap \Omega_{\text{pen}}$ , where  $\Omega_3^\delta$  and  $\Omega_4^\delta$  are respectively defined in (18) and (19). It comes from Lemma 8.16 and our assumption on  $\text{pen}(V)$  that  $P((\Omega_{\text{min pen}}^\delta)^c) \leq \epsilon + 2\delta^{-1}$ . Moreover, on  $\Omega_{\text{min pen}}^\delta$ , we have, from (23), for all  $\eta > 0$ ,

$$\begin{aligned} (1-\eta) \left\| P_{i|S} - P_{i|\hat{V}} \right\|_P^2 + r_1 \left\| \hat{P}_{i|\hat{V}} - P_{i|\hat{V}} \right\|_P^2 + (1+r_1) \left( \left\| \hat{P}_{i|\hat{V}} - P_{i|\hat{V}} \right\|_{\hat{P}}^2 - \left\| \hat{P}_{i|\hat{V}} - P_{i|\hat{V}} \right\|_P^2 \right) \\ \leq (1+\eta) \left\| P_{i|S} - P_{i|V} \right\|_P^2 + r_2 \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \\ + (1+r_2) \left( \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 - \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \right) + \left( \frac{17}{\eta} + 1 \right) \frac{\ln(N_s^2 \delta)}{3n}. \end{aligned}$$

Let  $C$  be the constant given by Lemma 8.8 and let

$$\Omega_* = \left\{ \forall V \in \mathcal{V}_s, \left| \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 - \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \right| \leq C\epsilon \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \right\}.$$

It comes from Lemma 8.8 that  $P(\Omega_*) \geq 1 - \delta^{-1}$ . Moreover, on  $\Omega_{\text{comp}} \cap \Omega_*$ , we have, from (23), for all  $0 < \eta < 1$ ,

$$\begin{aligned} (1-\eta) \left\| P_{i|S} - P_{i|\hat{V}} \right\|_P^2 + (r_1 - C(1+r_1)\epsilon) \left\| \hat{P}_{i|\hat{V}} - P_{i|\hat{V}} \right\|_P^2 \leq \\ \leq (1+\eta) \left\| P_{i|S} - P_{i|V} \right\|_P^2 + (r_2 + C(1+r_2)\epsilon) \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 + \frac{6 \ln(N_s^2 \delta)}{\eta n}. \quad \square \end{aligned}$$

## 7.7 Proof of Theorem 4.3:

Let  $\Omega_{\text{pen}}$  and  $\Omega_{\text{pen}}^{(2)}$  be the events, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,  $0 \leq \text{pen}(V) \leq (1-r)p_2(V)$  and for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,  $0 \leq \text{pen}(V) \leq (1-r)p_2(V)$ . It comes from (21) that, on  $\Omega_{\text{pen}}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$K(P_{i|S}, P_{i|\hat{V}}) - p_2(\hat{V}) \leq K(P_{i|S}, P_{i|V}) - rp_2(V) + L(V) - L(\hat{V}).$$

It comes from (21) that, on  $\Omega_{\text{pen}}^{(2)}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$K(P_{i|S}, P_{i|\hat{V}_{(2)}}) - p_2(\hat{V}_{(2)}) \leq K(P_{i|S}, P_{i|V}) - rp_2(V) + L(V) - L(\hat{V}_{(2)}).$$

Let  $\Omega_{\text{prob}}(\delta)$  be the event defined on Lemma 8.20 and  $\Omega_L(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}$ , for all  $\eta > 0$

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).$$

From Lemmas 8.20 and 8.23, we have  $P(\Omega_{\text{prob}}(\delta) \cap \Omega_L(\delta)) \geq 1 - 2\delta$  and, on  $\Omega_{\text{prob}}(\delta) \cap \Omega_L(\delta) \cap \Omega_{\text{pen}}$ , we have, for  $\eta = 1$ ,

$$-p_2(\hat{V}) \leq 2K(P_{i|S}, P_{i|V}) - rp_2(V) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2p_*} \right).$$

Let  $\Omega_L^{(2)}(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ , for all  $\eta > 0$

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).$$

From Lemmas 8.20 and 8.23, we have  $P(\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta)) \geq 1 - 2\delta$  and, on  $\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta) \cap \Omega_{pen}^{(2)}$ , we have, for  $\eta = 1$ ,

$$-p_2(\widehat{V}_{(2)}) \leq 2K(P_{i|S}, P_{i|V}) - r p_2(V) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2p_*} \right).$$

## 7.8 Proof of Theorem 4.4:

Let  $\Omega_{pen}$  be the event, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,  $(1 + r_1)p_2(V) \leq \text{pen}(V) \leq (1 + r_2)p_2(V)$ . It comes from (21) that, on  $\Omega_{pen}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$\begin{aligned} K(P_{i|S}, P_{i|\widehat{V}}) + r_1 p_1(\widehat{V}) + (1 + r_1)(p_2(\widehat{V}) - p_1(\widehat{V})) \\ \leq K(P_{i|S}, P_{i|V}) + r_2 p_1(V) + (1 + r_2)(p_2(V) - p_1(V)) + L(V) - L(\widehat{V}) \end{aligned}$$

Let  $\Omega_{prob}(\delta)$  be the event defined on Lemma 8.20 and  $\Omega_L(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}$ , for all  $\eta > 0$ ,

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).$$

From Lemmas 8.20 and 8.23, we have  $P(\Omega_{prob}(\delta) \cap \Omega_L(\delta)) \geq 1 - 2\delta$  and, on  $\Omega_{prob}(\delta) \cap \Omega_L(\delta) \cap \Omega_{pen}$ , we have, from Lemma 8.22, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$|p_1(V) - p_2(V)| \leq \frac{C}{\sqrt{\Lambda}} p_1(V).$$

We obtain that, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}$ ,

$$\begin{aligned} (1 - \eta)K(P_{i|S}, P_{i|\widehat{V}}) + \left( r_1 - \frac{C(1 + r_1)}{\sqrt{\Lambda}} \right) p_1(\widehat{V}) \\ \leq (1 + \eta)K(P_{i|S}, P_{i|V}) + \left( r_2 + \frac{C(1 + r_2)}{\sqrt{\Lambda}} \right) p_1(V) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right). \end{aligned}$$

Let  $\Omega_{pen}^{(2)}$  be the event, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,  $(1 + r_1)p_2(V) \leq \text{pen}(V) \leq (1 + r_2)p_2(V)$ . It comes from (21) that, on  $\Omega_{pen}^{(2)}$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$\begin{aligned} K(P_{i|S}, P_{i|\widehat{V}_{(2)}}) + r_1 p_1(\widehat{V}_{(2)}) + (1 + r_1)(p_2(\widehat{V}_{(2)}) - p_1(\widehat{V}_{(2)})) \\ \leq K(P_{i|S}, P_{i|V}) + r_2 p_1(V) + (1 + r_2)(p_2(V) - p_1(V)) + L(V) - L(\widehat{V}_{(2)}) \end{aligned}$$

Let  $\Omega_{prob}(\delta)$  be the event defined on Lemma 8.20 and  $\Omega_L^{(2)}(\delta)$  be the event, for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ , for all  $\eta > 0$ ,

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).$$

From Lemmas 8.20 and 8.23, we have  $P(\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta)) \geq 1 - 2\delta$  and, on  $\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta) \cap \Omega_{pen}$ , we have, from Lemma 8.22, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$|p_1(V) - p_2(V)| \leq \frac{C}{\sqrt{\Lambda}} p_1(V).$$

We obtain that, for all  $V$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ ,

$$\begin{aligned} (1 - \eta)K(P_{i|S}, P_{i|\widehat{V}_{(2)}}) + \left(r_1 - \frac{C(1 + r_1)}{\sqrt{\Lambda}}\right) p_1(\widehat{V}_{(2)}) \\ \leq (1 + \eta)K(P_{i|S}, P_{i|V}) + \left(r_2 + \frac{C(1 + r_2)}{\sqrt{\Lambda}}\right) p_1(V) + \frac{\ln(N_s^2 \delta)}{n} \left(4 \ln n + \frac{3}{2\eta p_*}\right). \end{aligned}$$

## 8 Probabilistic Tools

**Lemma 8.1.** *Let  $x$  in  $\mathcal{X}(S)$ , let  $V$  be a finite subset of  $S$  and let  $Q, R$  be two probability measures on  $\mathcal{X}(V)$  such that  $R(x(V/\{i\})) > 0$ . We have*

$$Q_{i|V}(x) - R_{i|V}(x) = \frac{Q(x(V)) - R(x(V)) + Q_{i|V}(x)(R(x(V/\{i\})) - Q(x(V/\{i\})))}{R(x(V/\{i\}))}.$$

The lemma immediately follows from the fact that  $Q_{i|V}(x)Q(x(V/\{i\})) = Q(x(V))$  and  $R_{i|V}(x) = R(x(V))/R(x(V/\{i\}))$ .

We recall the bound given by Bousquet [Bou02] for the deviation of the supremum of the empirical process.

**Theorem 8.2.** *Let  $X_1, \dots, X_n$  be i.i.d. random variables valued in a measurable space  $(A, \mathcal{X})$ . Let  $\mathcal{F}$  be a class of real valued functions, defined on  $A$  and bounded by  $b$ . Let  $v^2 = \sup_{f \in \mathcal{F}} P[(f - Pf)^2]$  and  $Z = \sup_{f \in \mathcal{F}} (P_n - P)f$ . Then, for all  $x > 0$ ,*

$$P\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n}\right) \leq e^{-x}. \quad (24)$$

Bousquet's result is a generalization of the elementary Benett's inequality.

**Theorem 8.3.** *Let  $X_1, \dots, X_n$  be i.i.d. random variables, real valued and bounded by  $b$ . Let  $v^2 = \text{Var}(X_1)$  and  $\bar{X}_n = n^{-1} \sum_{i=1}^n (X_i - \mathbb{E}(X))$ . Then, for all  $x > 0$ ,*

$$P\left(\bar{X}_n > \sqrt{\frac{2v^2 x}{n}} + \frac{bx}{3n}\right) \leq e^{-x}. \quad (25)$$

Let us recall some well known tools of empirical processes theory.

**Definition 8.4.** *The covering number  $N(\epsilon, T, d)$  is the minimal number of balls of radius  $\epsilon$  with centers in  $T$  needed to cover  $T$ . The entropy is the log of the covering number  $H(\epsilon, T, d) = \log(N(\epsilon, T, d))$ .*

**Definition 8.5.** *An  $\epsilon$ -separated subset of  $T$  is a subset  $\{t_k\}$  of elements of  $T$  whose pairwise distance is strictly larger than  $\epsilon$ . The packing number  $M(\epsilon, T, d)$  is the maximum size of an  $\epsilon$ -separated subset of  $T$ .*

Those quantities are related by the famous following lemma.

**Lemma 8.6.** *(Kolmogorov and Tikhomirov [KT63]) Let  $(T, d)$  be a metric space and let  $\epsilon > 0$ ,*

$$N(\epsilon, T, d) \leq M(\epsilon, T, d) \leq N(\epsilon/2, T, d).$$



## 8.1 Concentration for Slope with quadratic risk

The aim of this section is to prove the following result.

**Theorem 8.7.** *Let  $(S, A, P)$  be a random field and let  $V$  be a subspace in  $\mathcal{V}_s$ . Let  $\mathcal{X}'(V) = \{x \in \mathcal{X}(V), P(x(V)) \neq 0\}$  and let  $p_-^V = \inf_{x \in \mathcal{X}'(V)} P(x(V))$ .*

*Let  $Z = \sup_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V)) - P(x(V))|}{P(x(V))}$ . For all  $\delta > 1$ , with probability larger than  $1 - \delta^{-1}$ ,*

$$Z \leq \frac{64\sqrt{2}}{\sqrt{np_-^V}} \sqrt{\ln\left(\frac{16}{p_-^V}\right)} + \frac{2048}{np_-^V} \ln\left(\frac{16}{p_-^V}\right) + \sqrt{\frac{2\ln(\delta)}{np_-^V}} + 2\frac{\ln(\delta)}{np_-^V}.$$

Let us state an important consequence of Theorem 8.7.

**Lemma 8.8.** *Assume that  $\inf_{V \in \mathcal{V}_s} p_-^V \geq \varepsilon^{-2} n^{-1} \ln(nN_s\delta)$ . There exists an absolute constant  $C$  such that, with probability larger than  $1 - \delta^{-1}$ , for all  $V$  in  $\mathcal{V}_s$ ,*

$$\left| \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 - \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 \right| \leq C\varepsilon \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2.$$

**Proof:** Let  $V$  in  $\mathcal{V}_s$  and let  $\mathcal{X}'(V) = \{x \in \mathcal{X}(V), P(x(V/\{i\})) \neq 0\}$ . We have

$$\begin{aligned} & \left| \left\| \hat{P}_{i|V} - P_{i|V} \right\|_{\hat{P}}^2 - \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2 \right| \\ & \leq \sum_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V/\{i\})) - P(x(V/\{i\}))|}{a} \left( \hat{P}_{i|V}(x) - P_{i|V}(x) \right)^2 \\ & \leq \sup_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V/\{i\})) - P(x(V/\{i\}))|}{P(x(V/\{i\}))} \left\| \hat{P}_{i|V} - P_{i|V} \right\|_P^2. \end{aligned}$$

We take a union bound in Theorem 8.7 and we obtain, since  $\inf_{V \in \mathcal{V}_s} p_-^V \geq \varepsilon^{-2} n^{-1} \ln(nN_s\delta)$  that there exists an absolute constant  $C$  such that

$$\forall V \in \mathcal{V}_s, \sup_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V/\{i\})) - P(x(V/\{i\}))|}{P(x(V/\{i\}))} \leq C\varepsilon.$$

In the remainder of this section, we prove Theorem 8.7.

**Proposition 8.9.** *Let  $P$  be a probability measure on  $\mathcal{X}(S)$  and let  $V$  be a finite subset of  $S$ . Let  $\mathcal{X}'(V) = \{x \in \mathcal{X}(V), P(x(V)) \neq 0\}$  and let  $p_-^V = \inf_{x \in \mathcal{X}'(V)} P(x(V))$ .*

*Let  $Z = \sup_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V)) - P(x(V))|}{P(x(V))}$ . For all  $\delta > 0$ , with probability larger than  $1 - \delta^{-1}$ ,*

$$Z \leq 2\mathbb{E}(Z) + \sqrt{\frac{2\ln(\delta)}{np_-^V}} + 2\frac{\ln(\delta)}{np_-^V}.$$

Proposition 8.9 is a straightforward consequence of Bousquet's version of Talagrand's inequality, that we apply to the class of functions  $\mathcal{F} = \{(P(x(V)))^{-1} 1_{x(V)}\}$ .

The second proposition let us compute this expectation.

**Proposition 8.10.** *Let  $P$  be a probability measure on  $\mathcal{X}(S)$  and let  $V$  be a finite subset of  $S$ . Let  $\mathcal{X}'(V) = \{x \in \mathcal{X}(V), P(x(V)) \neq 0\}$  and let  $p_-^V = \inf_{x \in \mathcal{X}'(V)} P(x(V))$ .*

$$\mathbb{E} \left( \sup_{x \in \mathcal{X}'(V)} \frac{|\hat{P}(x(V)) - P(x(V))|}{P(x(V))} \right) \leq \frac{32\sqrt{2}}{\sqrt{np_-^V}} \sqrt{\ln\left(\frac{16}{p_-^V}\right)} + \frac{1024}{np_-^V} \ln\left(\frac{16}{p_-^V}\right).$$

Proposition 8.10 was proved in [LT11]. We recall the proof here for the sake of completeness. Let  $(A_i)_{i \in I}$  be a collection of sets such that, for all  $i \neq j \in I$ ,  $A_i \cap A_j = \emptyset$  and let  $(\alpha_i)_{i \in I}$  be a collection of positive real numbers. Let  $Z_I = \sup_{t \in \mathcal{F}_I} |(P_n - P)t|$ , where  $\mathcal{F}_I = \{t_i = \alpha_i 1_{A_i}\}$ . Here and in the rest of the proof,  $\alpha^* = \sup_{i \in I} \alpha_i$ ,  $p_* = \sup_{i \in I} \alpha_i^2 P(A_i)$ . The following result can be derived from classical chaining arguments (see for example [Bou02]).

**Lemma 8.11.** *Let  $\mathcal{F}$  be a class of functions, then*

$$\mathbb{E} \left( \sup_{t \in \mathcal{F}} |(P_n - P)t| \right) \leq \frac{16\sqrt{2}}{\sqrt{n}} \mathbb{E} \left( \int_0^{D_n/2} H^{1/2}(u, \mathcal{F}, d_{2,P_n}) du \right),$$

where the distance  $d_{2,P_n}(t, t') = \sqrt{P_n[(t - t')^2]}$  and the diameter  $D_n = \sqrt{\sup_{t \in \mathcal{F}} P_n(t^2)}$ .

In order to apply this result to  $\mathcal{F} = \mathcal{F}_I$ , we compute the entropy of  $\mathcal{F}_I$ . For all  $i \neq j$ , since  $A_i \cap A_j = \emptyset$ ,

$$(t_i - t_j)^2 = (\alpha_i 1_{A_i} - \alpha_j 1_{A_j})^2 = \alpha_i^2 1_{A_i} + \alpha_j^2 1_{A_j}.$$

Hence  $d_{2,P_n}(t_i, t_j) = \sqrt{\alpha_i^2 P_n(A_i) + \alpha_j^2 P_n(A_j)}$ .

Consider an  $\epsilon$ -separated set  $T_\epsilon = \{t_{i_1}, \dots, t_{i_N}\}$  in  $(\mathcal{F}_I, d_{2,P_n})$ , it comes from the previous computation that, for all  $k \neq k'$ ,

$$\alpha_{i_k}^2 P_n(A_{i_k}) + \alpha_{i_{k'}}^2 P_n(A_{i_{k'}}) \geq \epsilon^2.$$

Hence, there is at least  $N - 1$  indexes  $k \in \{1, \dots, N\}$  such that

$$\alpha_{i_k}^2 P_n(A_{i_k}) \geq \frac{\epsilon^2}{2}.$$

It follows that

$$1 = \sum_{i \in I} P_n(A_i) \geq \sum_{k=1}^N P_n(A_{i_k}) \geq \frac{\epsilon^2(N-1)}{2(\alpha^*)^2}.$$

Hence  $N \leq 1 + 2(\alpha^*)^2 \epsilon^{-2}$ , thus

$$H(\epsilon, \mathcal{F}_I, d_{2,P_n}) \leq \log(1 + 2(\alpha^*)^2 \epsilon^{-2}).$$

We deduce from this inequality and Lemma 8.11 that

$$\begin{aligned} \mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right) &\leq \frac{16\sqrt{2}}{\sqrt{n}} \mathbb{E} \left( \int_0^{\sqrt{\hat{p}_n^*}/2} \sqrt{\log(1 + 2(\alpha^*)^2 \epsilon^{-2})} d\epsilon \right) \\ &\leq \frac{32}{\sqrt{n}} \mathbb{E} \left( \int_0^{\sqrt{\hat{p}_n^*}/2} \sqrt{\log(2\alpha^* \epsilon^{-1})} d\epsilon \right), \end{aligned}$$

where  $\hat{p}_n^* = \sup_{i \in I} \alpha_i^2 P_n(A_i)$ . Now, let us recall the following elementary lemma.

**Lemma 8.12.** *For all positive  $K, A$  such that  $K/A > e$ , we have*

$$\int_0^A \sqrt{\log(Kx^{-1})} dx \leq 2A \sqrt{\log\left(\frac{K}{A}\right)}$$

Actually,

$$\int_0^A \sqrt{\log(Kx^{-1})} dx = K \int_{K/A}^\infty \frac{\sqrt{\log(x)}}{x^2} dx = A \sqrt{\log\left(\frac{K}{A}\right)} + \frac{K}{2} \int_{K/A}^\infty \frac{1}{u^2 \sqrt{\log u}} du.$$

Since  $K/A > e$ ,  $\frac{1}{u^2 \sqrt{\log u}} \leq \frac{\sqrt{\log u}}{u^2}$  on  $[K/A, \infty[$ . The result follows.

By definition,  $\hat{p}_n \leq (\alpha^*)^2$ , hence  $2\alpha^*/(\sqrt{\hat{p}_n^*}/2) \geq 4 > e$ , we deduce from Lemma 8.12 that

$$\mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right) \leq \frac{32}{\sqrt{n}} \mathbb{E} \left[ \sqrt{\hat{p}_n^*} \sqrt{\log \left( \frac{4\alpha^*}{\sqrt{\hat{p}_n^*}} \right)} \right].$$

Let us now give another simple lemma.

**Lemma 8.13.** *The function  $f : x \mapsto x\sqrt{\log(K/x)}$ , defined on  $(0, K)$  is positive, non decreasing on  $(0, K/e^{1/2})$  and strictly concave.*

The proof of the lemma is straightforward from the computations

$$f'(x) = \sqrt{\log(K/x)} - \frac{1}{2\sqrt{\log(K/x)}}, \quad f''(x) = -\frac{1}{2x\sqrt{\log(K/x)}} - \frac{1}{4x(\sqrt{\log(K/x)})^3}.$$

It comes from Lemma 8.13 and Jensen's inequality that

$$\mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right) \leq \frac{32}{\sqrt{n}} \mathbb{E} \left( \sqrt{\hat{p}_n^*} \right) \sqrt{\log \left( \frac{4\alpha^*}{\mathbb{E}(\sqrt{\hat{p}_n^*})} \right)}.$$

Now it comes from Jensen inequality that

$$\mathbb{E} \left[ \sqrt{\hat{p}_n^*} \right] \leq \sqrt{\mathbb{E}[\hat{p}_n^*]} \leq \sqrt{p^*} + \sqrt{\alpha^* \mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right)}.$$

It is clear from its definition that  $p^* \leq (\alpha^*)^2$ . Moreover, as  $P_n$  and  $P$  are probability measures, we have, for all  $t$  in  $\mathcal{F}_I$ ,  $|(P_n - P)t| \leq 2\alpha^*$ . Hence,  $\sqrt{\alpha^* \mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right)} \leq \sqrt{2}\alpha^*$ . We deduce from these inequalities that

$$\sqrt{p^*} + \sqrt{\alpha^* \mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right)} \leq (1 + \sqrt{2})\alpha^* \leq (4\alpha^*)/e^{1/2}.$$

Hence, it comes from Lemma 8.13 that, if  $E = \mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right)$

$$\begin{aligned} E &\leq \frac{32}{\sqrt{n}} \left( \sqrt{p^*} + \sqrt{\alpha^* E} \right) \sqrt{\log \left( \frac{4\alpha^*}{\sqrt{p^*} + \sqrt{\alpha^* E}} \right)} \\ &\leq \frac{32}{\sqrt{n}} \left( \sqrt{p^*} + \sqrt{\alpha^* E} \right) \sqrt{\log \left( \frac{4\alpha^*}{\sqrt{p^*}} \right)} \end{aligned}$$

It is then straightforward that

$$\mathbb{E} \left( \sup_{t \in \mathcal{F}_I} |(P_n - P)t| \right) \leq \frac{64}{\sqrt{n}} \sqrt{p^* \log \left( \frac{4\alpha^*}{\sqrt{p^*}} \right)} + \frac{2048}{n} \alpha^* \log \left( \frac{4\alpha^*}{\sqrt{p^*}} \right). \quad (26)$$

In order to conclude the proof of Proposition 8.10, for all  $x \in \mathcal{X}'(V)$ , let  $A_x = x(V)$ ,  $\alpha_x = [P(x(V))]^{-1}$ . We have

$$\alpha^* = \sup_{x \in \mathcal{X}'(V)} [P(x(V))]^{-1} = \frac{1}{p_-^V}, \quad p^* = \sup_{x \in \mathcal{X}'(V)} [P(x(V))]^{-2} P(x(V)) \leq \frac{1}{p_-^V}.$$

Therefore, the proposition is straightforward from inequality (26).

## 8.2 Concentration of the variance term in quadratic risk

The aim of this section is to prove the following concentration result, that is at the center of the main proofs.

**Theorem 8.14.** *Let  $V$  be a finite subset of  $S$ . Let  $b > 0$  and let  $\mathcal{X}^b(V) = \{x \in \mathcal{X}(V), P(x(V)) \geq b^{-2}\}$ . For all  $x > 0$ ,  $\eta > 0$ , we have,*

$$P \left( \sum_{x \in \mathcal{X}^b(V)} \frac{(\hat{P}(x(V)) - P(x(V)))^2}{P(x(V))} \leq (1 + \eta)^3 \frac{a^v}{n} + \frac{4x}{\eta n} + \frac{32b^2 x^2}{\eta^3 n^2} \right) \geq 1 - e^{-x}.$$

**Proof:** Let us first recall the following consequence of Cauchy-Schwarz inequality.

**Lemma 8.15.** *Let  $I$  be a finite set and let  $(b_i)_{i \in I}$  be a collection of real numbers. We have*

$$\sum_{i \in I} b_i^2 = \left( \sup_{(a_i)_{i \in I}, \sum_{i \in I} a_i^2 \leq 1} \sum_{i \in I} a_i b_i \right)^2.$$

**Proof:** The lemma is obviously satisfied if all the  $b_i = 0$ . Assume now that it is not the case. By Cauchy Schwarz inequality, we have, for all collection  $(a_i)_{i \in I}$  such that  $\sum_{i \in I} a_i^2 \leq 1$ ,

$$\left( \sum_{i \in I} a_i b_i \right)^2 \leq \sum_{i \in I} a_i^2 \sum_{i \in I} b_i^2 \leq \sum_{i \in I} b_i^2.$$

Moreover, consider for all  $i$  in  $I$ ,  $a_i = b_i / \sqrt{\sum_{i \in I} b_i^2}$ , we have  $\sum_{i \in I} a_i^2 = 1$  and  $\sum_{i \in I} a_i b_i = \sqrt{\sum_{i \in I} b_i^2}$ , which concludes the proof.

Let us now introduce the following set.

$$B_V^b = \left\{ f : \mathcal{X}^b(V) \rightarrow \mathbb{R} \text{ such that } f = \sum_{x \in \mathcal{X}^b(V)} \frac{\alpha_x \mathbf{1}_x}{\sqrt{P(x(V))}}, \text{ where } \sum_{x \in \mathcal{X}^b(V)} \alpha_x^2 \leq 1. \right\}.$$

Let  $P$  and  $P_n$  be the following operators, defined for all functions  $f$ , by  $P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$  and, for all functions  $f$  in  $L^1(P)$ , by  $Pf = \int f(x) dP(x)$ . Using Lemma 8.15 with  $I = \mathcal{X}^b(V)$  and

$$b_x = \frac{\hat{P}(x(V)) - P(x(V))}{\sqrt{P(x(V))}} = (P_n - P) \left( \frac{\mathbf{1}_{x(V)}}{\sqrt{P(x(V))}} \right),$$

we obtain,

$$\sum_{x \in \mathcal{X}^b(V)} \frac{(\hat{P}(x(V)) - P(x(V)))^2}{P(x(V))} = \left( \sup_{f \in B_V^b} (P_n - P)f \right)^2,$$

The functions  $f$  in  $B_V^p$  satisfy

$$\text{Var}(f(X)) \leq P f^2 = \sum_{x \in \mathcal{X}^b(V)} \frac{\alpha_x^2}{P(x(V))} P(x(V)) \leq 1, \quad \|f\|_\infty \leq \sup_{x \in \mathcal{X}^b(V)} \frac{1}{\sqrt{P(x(V))}} \leq b.$$

From Theorem 8.2, we have then, for all  $\eta > 0$ , for all  $x > 0$ ,

$$P \left( \sup_{f \in B_V^b} (P_n - P)f > (1 + \eta) \mathbb{E} \left( \sup_{f \in B_V^b} (P_n - P)f \right) + \sqrt{\frac{2x}{n}} + \left( \frac{1}{3} + \frac{1}{\eta} \right) \frac{bx}{n} \right) \leq e^{-x}.$$

From Cauchy-Schwarz inequality, we have then

$$\begin{aligned} \mathbb{E} \left( \sup_{f \in B_V^b} (P_n - P)f \right) &\leq \sqrt{\mathbb{E} \left( \left( \sup_{f \in B_V^b} (P_n - P)f \right)^2 \right)} \\ &= \sqrt{\sum_{x \in \mathcal{X}(V), P(x(V)) \neq 0} \frac{\mathbb{E} \left( \left( \hat{P}(x(V)) - P(x(V)) \right)^2 \right)}{P(x(V/\{i\}))}} \\ &= \sqrt{\sum_{x \in \mathcal{X}(V), P(x(V)) \neq 0} \frac{\text{Var}(1_{X(V)=x(V)})}{nP(x(V/\{i\}))}} \leq \sqrt{\frac{a^v}{n}}. \end{aligned}$$

We have obtain that

$$P \left( \sup_{f \in B_V^b} (P_n - P)f > (1 + \eta) \sqrt{\frac{a^v}{n}} + \sqrt{\frac{2x}{n}} + \left( \frac{1}{3} + \frac{1}{\eta} \right) \frac{bx}{n} \right) \leq e^{-x}.$$

Since  $B_V^b$  is symmetric,  $\sup_{f \in B_V^b} (P_n - P)f \geq 0$ . We can therefore take the square in the previous inequality to conclude the proof of the Theorem.

### 8.3 Concentration of the remainder term in the quadratic case

Let us now give some important concentration inequalities.

**Lemma 8.16.** *Let  $V, V'$  be two subsets in  $\mathcal{V}_s$ . For all  $\delta > 0$ , we have, with probability larger than  $1 - \delta$ ,*

$$\begin{aligned} \frac{1}{a} \sum_{x \in \mathcal{X}((V \cup V'))} (\hat{P}(x((V \cup V')/\{i\})) - P(x((V \cup V')/\{i\}))) (P_{i|V}^2(x) - P_{i|V'}^2(x)) \\ \leq 2 \|P_{i|V} - P_{i|V'}\|_P \sqrt{2 \frac{\ln(\delta)}{n}} + \frac{\ln(\delta)}{3n}. \end{aligned} \quad (27)$$

$$\begin{aligned} \frac{1}{a} \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) (P_{i|V'}(x) - P_{i|V}(x)) \\ \leq \|P_{i|V} - P_{i|V'}\|_P \sqrt{2 \frac{\ln(\delta)}{n}} + \frac{\ln(\delta)}{3n}. \end{aligned} \quad (28)$$

**Proof of Lemma 8.16:** Let  $f_1$  be the real valued function defined on  $\mathcal{X}((V \cup V')/\{i\})$  by

$$f_1 = \sum_{x \in \mathcal{X}((V \cup V'))} \left( P_{i|V}^2(x) - P_{i|V'}^2(x) \right) 1_{x((V \cup V')/\{i\})}.$$

$$f_2 = \sum_{x \in \mathcal{X}((V \cup V'))} \left( P_{i|V}(x) - P_{i|V'}(x) \right) 1_{x(V \cup V')}.$$

We have

$$f_1 = \sum_{x \in \mathcal{X}((V \cup V')/\{i\})} 1_{x((V \cup V')/\{i\})} \sum_{b \in A} \left( P_{i|V}^2(x_b) - P_{i|V'}^2(x_b) \right).$$

$f_1$  is upper bounded by  $\max_{x \in \mathcal{X}((V \cup V')/\{i\})} \sum_{b \in A} \left( P_{i|V}^2(x_b) - P_{i|V'}^2(x_b) \right) \leq a$ .  $f_2$  is upper bounded by  $\max_{x \in \mathcal{X}(V \cup V')} |P_{i|V}(x) - P_{i|V'}(x)|$ . Since, for all  $x \neq x'$  in  $\mathcal{X}((V \cup V')/\{i\})$ ,  $1_{x((V \cup V')/\{i\})} 1_{x'((V \cup V')/\{i\})} = 0$ , we have

$$\begin{aligned} \text{Var}(f_1(X)) &= \sum_{x \in \mathcal{X}((V \cup V')/\{i\})} P(x((V \cup V')/\{i\})) \left( \sum_{b \in A} P_{i|V}^2(x_b) - P_{i|V'}^2(x_b) \right)^2 \\ &\leq a \sum_{x \in \mathcal{X}((V \cup V')/\{i\})} P(x((V \cup V')/\{i\})) \sum_{b \in A} \left( P_{i|V}^2(x_b) - P_{i|V'}^2(x_b) \right)^2 \\ &\leq 4a \sum_{x \in \mathcal{X}(V \cup V')} \left( P_{i|V}(x) - P_{i|V'}(x) \right)^2 P(x(V \cup V')/\{i\}) \\ &= 4a^2 \|P_{i|V} - P_{i|V'}\|_P^2. \end{aligned}$$

Since, for all  $x \neq x'$  in  $\mathcal{X}(V \cup V')$ ,  $1_{x(V \cup V')} 1_{x'(V \cup V')} = 0$ , we have

$$\text{Var}(f_2(X)) \leq \sum_{x \in \mathcal{X}((V \cup V'))} \left( P_{i|V}(x) - P_{i|V'}(x) \right)^2 P(V \cup V') \leq a \|P_{i|V} - P_{i|V'}\|_P^2.$$

Inequality 27 is therefore a consequence of Benett's inequality, see Theorem 8.3. We obtain Inequality 28 exactly with the same arguments.  $\square$

## Pythagoras relation

Let us give here Pythagoras relation that we used several times.

**Proposition 8.17.** *Let  $(S, A, P)$  be a random field, let  $i$  in  $S$  and let  $V$  be a subset of  $S$  and let  $f$  be a function defined on  $\mathcal{X}(V)$ . Then, the following relations hold*

$$\begin{aligned} \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|S}(x) dP(x(S/\{i\})) &= \int_{x \in \mathcal{X}(V)} f(x(V)) P_{i|V}(x) dP(x(V/\{i\})) \\ &= \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|V}(x) dP(x(S/\{i\})). \end{aligned}$$

In particular, we have

$$\begin{aligned} \|\hat{P}_{i|V} - P_{i|S}\|_P^2 &= \|\hat{P}_{i|V} - P_{i|V}\|_P^2 + \|P_{i|V} - P_{i|S}\|_P^2. \\ \|P_{i|V} - P_{i|S}\|_P^2 - \|P_{i|S}\|_P^2 &= -\|P_{i|V}\|_P^2 \end{aligned}$$

**Proof:** The first inequality comes from the following computations. For all  $x$  in  $\mathcal{X}(V)$  and  $y$  in  $\mathcal{X}(S/V)$ , let  $x(V) \oplus y(S/V)$  be the configuration on  $\mathcal{X}(S)$  such that  $(x(V) \oplus y(S/V))(j) = x(j)$  for all  $j$  in  $V$  and  $(x(V) \oplus y(S/V))(j) = y(j)$  for all  $j$  in  $S/V$ . By definition of the conditional probabilities  $P_{i|V}(x)$ , we have

$$\begin{aligned} & \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|S}(x) dP(x(S/\{i\})) \\ &= \int_{x \in \mathcal{X}(V)} f(x(V)) dP(x(V/\{i\})) \int_{y \in \mathcal{X}(S/V)} P_{i|S}(x(V) \oplus y(S/V)) dP(y(S/V)|x(V)) \\ &= \int_{x \in \mathcal{X}(V)} f(x(V)) P_{i|V}(x) dP(x(V/\{i\})). \end{aligned}$$

The second inequality is a straightforward consequence of the first one. For the third one, we apply the second inequality to  $f(x(V)) = \hat{P}_{i|V} - P_{i|V}$ , we have

$$\int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|S}(x) dP(x(S/\{i\})) = \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|V}(x) dP(x(S/\{i\})).$$

Thus,

$$\begin{aligned} \|\hat{P}_{i|V} - P_{i|S}\|_P^2 &= \|f(x(V)) + P_{i|V} - P_{i|S}\|_P^2 = \|f(x(V))\|_P^2 + \|P_{i|V} - P_{i|S}\|_P^2 + \\ & \quad \frac{2}{a} \left( \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|V}(x) dP(x(S/\{i\})) - \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|S}(x) dP(x(S/\{i\})) \right) \\ &= \|\hat{P}_{i|V} - P_{i|V}\|_P^2 + \|P_{i|V} - P_{i|S}\|_P^2. \end{aligned}$$

For the last inequality, we use the second one with  $f(x(V)) = P_{i|V}(x)$ , we have

$$\begin{aligned} \|P_{i|V} - P_{i|S}\|_P^2 &= \|P_{i|V}\|_P^2 + \|P_{i|S}\|_P^2 - \frac{2}{a} \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|S}(x) dP(x(S/\{i\})) \\ &= \|P_{i|V}\|_P^2 + \|P_{i|S}\|_P^2 - \frac{2}{a} \int_{x \in \mathcal{X}(S)} f(x(V)) P_{i|V}(x) dP(x(S/\{i\})) \\ &= \|P_{i|V}\|_P^2 + \|P_{i|S}\|_P^2 - 2\|P_{i|V}\|_P^2 = \|P_{i|S}\|_P^2 - \|P_{i|V}\|_P^2. \end{aligned}$$

## 8.4 Basic tools for Küllback Loss

Let  $s$  be an integer larger than  $e$ . Let  $\mathcal{V}_s$  be the collection of subsets of  $V$  with cardinality smaller than  $s$ . Let  $N_s$  be the cardinality of  $\mathcal{V}_s$ . Let  $i$  be a site in  $V$ . Let us first give an elementary lemma on Küllback losses. It is a slightly sharper version of Lemma 6.3 in [CT06b].

**Lemma 8.18.** *Let  $P, Q$  be two probability measures on a finite space  $A$  such that, for all  $a$  in  $A$ ,  $|P(a) - Q(a)| \leq \eta Q(a)$ , with  $\eta \leq 1/3$ . Then*

$$\left( \frac{1}{2} - \frac{7\eta}{6} \right) \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)} \leq \sum_{a \in A} P(a) \ln \left( \frac{P(a)}{Q(a)} \right) \leq \left( \frac{1}{2} + \frac{5\eta}{6} \right) \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)}.$$

**Proof:** Let us first prove the following inequality, that is valid for all  $x \leq 1/3$ .

$$x - x^2 \left( \frac{1}{2} + \frac{\eta}{2} \right) \leq \ln(1+x) \leq x - x^2 \left( \frac{1}{2} - \frac{\eta}{2} \right).$$

It comes from the Taylor expansion.

$$\ln(1+x) = x - \frac{x^2}{2} + \sum_{k \geq 3} \frac{(-1)^{k+1} x^k}{k} \leq x - \frac{x^2}{2} + \frac{x^2 \eta}{3} \sum_{k \geq 0} \eta^k = x - x^2 \left( \frac{1}{2} - \frac{\eta}{3(1-\eta)} \right).$$

$$\ln(1+x) = x - \frac{x^2}{2} + \sum_{k \geq 3} \frac{(-1)^{k+1} x^k}{k} \geq x - \frac{x^2}{2} - \frac{x^2 \eta}{3} \sum_{k \geq 0} \eta^k = x - x^2 \left( \frac{1}{2} + \frac{\eta}{3(1-\eta)} \right).$$

We deduce from this inequality and the equality

$$\sum_{a \in A} P(a) \frac{P(a) - Q(a)}{Q(a)} = \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)}$$

that

$$\begin{aligned} \sum_{a \in A} P(a) \ln \left( \frac{P(a)}{Q(a)} \right) &\leq \sum_{a \in A} P(a) \frac{P(a) - Q(a)}{Q(a)} - \left( \frac{1}{2} - \frac{\eta}{2} \right) \sum_{a \in A} \frac{P(a)}{Q(a)} \frac{(P(a) - Q(a))^2}{Q(a)} \\ &= \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)} \left( \frac{1}{2} + \frac{\eta}{2} + \left| \frac{P(a)}{Q(a)} - 1 \right| \left( \frac{1}{2} - \frac{\eta}{2} \right) \right) \\ \sum_{a \in A} P(a) \ln \left( \frac{P(a)}{Q(a)} \right) &\geq \sum_{a \in A} P(a) \frac{P(a) - Q(a)}{Q(a)} - \left( \frac{1}{2} + \frac{\eta}{2} \right) \sum_{a \in A} \frac{P(a)}{Q(a)} \frac{(P(a) - Q(a))^2}{Q(a)} \\ &= \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)} \left( \frac{1}{2} - \frac{\eta}{2} - \left| \frac{P(a)}{Q(a)} - 1 \right| \left( \frac{1}{2} + \frac{\eta}{2} \right) \right) \end{aligned}$$

## 8.5 Basic Concentration Inequality

Let us now give an elementary concentration results derived from Benett's inequality.

**Lemma 8.19.** *Let  $\delta > 1$ . With probability larger than  $1 - \delta^{-1}$ , for all  $(V \times x) \in (\mathcal{V}_s \times \mathcal{X}(S))$ , we have*

$$\left| P(x(V)) - \hat{P}(x(V)) \right| \leq \sqrt{\frac{2P(x(V)) \ln(2a^s N_s \delta)}{n}} + \frac{\ln(2a^s N_s \delta)}{3n}.$$

**Proof:** Let  $V$  in  $\mathcal{V}_s$  and  $x$  in  $\mathcal{X}(V)$ , we have from Benett's inequality, for all  $t > 0$ ,

$$P \left( \left| P(x(V)) - \hat{P}(x(V)) \right| > \sqrt{\frac{2\text{Var}(\mathbf{1}_{x(V)})t}{n}} + \frac{t}{3n} \right) \leq 2e^{-t}.$$

We have  $\text{Var}(\mathbf{1}_{x(V)}) \leq P(x(V))$ . Hence, we conclude the proof with a union bound.

We deduce from Lemma 8.19 the following typicality results.

**Lemma 8.20.** *Let  $\Lambda \geq 100$ . Let  $\Omega_{\text{prob}}(\delta)$  be the following event,*

$$\left\{ \forall (V, x) \in \mathcal{V}_n \times \mathcal{X}(S), \left| P(x(V)) - \hat{P}(x(V)) \right| \leq \sqrt{\frac{2P(x(V)) \ln(2a^s N_s \delta)}{n}} + \frac{\ln(2a^s N_s \delta)}{3n} \right\}.$$

*We have  $P(\Omega_{\text{prob}}(\delta)) \geq 1 - \delta^{-1}$  and, on  $\Omega_{\text{prob}}(\delta)$ , for all  $V$  in  $\mathcal{V}_s$  and all  $x$  in  $\mathcal{X}(S)$  such that*

$$P(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n},$$



We have

$$\begin{aligned} \left| P(x(V)) - \hat{P}(x(V)) \right| &\leq 2\sqrt{\frac{P(x(V)) \ln(2a^s N_s \delta)}{n}} \leq \frac{2P(x(V))}{\sqrt{\Lambda}}. \\ \left| P_{i|V}(x) - \hat{P}_{i|V}(x) \right| &\leq \sqrt{\frac{11}{\Lambda}} P_{i|V}(x). \end{aligned}$$

On  $\Omega_{\text{prob}}(\delta)$ , for all  $V$  in  $\mathcal{V}_s$  and all  $x$  in  $\mathcal{X}(S)$  such that

$$\hat{P}(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n},$$

We have

$$\begin{aligned} \left| P(x(V)) - \hat{P}(x(V)) \right| &\leq 2\sqrt{\frac{P(x(V)) \ln(2a^s N_s \delta)}{n}} \leq \frac{2P(x(V))}{\sqrt{\Lambda}}. \\ \left| P_{i|V}(x) - \hat{P}_{i|V}(x) \right| &\leq \sqrt{\frac{11}{\Lambda}} P_{i|V}(x). \end{aligned}$$

**Proof:** When  $P(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n}$ , we have

$$\frac{\ln(2a^s N_s \delta)}{3n} \leq \frac{1}{3\sqrt{2\Lambda}} \sqrt{\frac{2P(x(V)) \ln(2a^s N_s \delta)}{n}}, \text{ and } \sqrt{\frac{P(x(V)) \ln(2a^s N_s \delta)}{n}} \leq \frac{P(x(V))}{\sqrt{\Lambda}}.$$

This gives the first inequalities, as  $\sqrt{2} + (3\sqrt{2\Lambda})^{-1} \leq 2$ . We also have, since  $P(x(V/\{i\})) \geq P(x(V))$ ,

$$\left| P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right| \leq 2\sqrt{\frac{P(x(V/\{i\})) \ln(2a^s N_s \delta)}{n}} \leq \frac{2P(x(V/\{i\}))}{\sqrt{\Lambda}}.$$

From Lemma 8.1, we have

$$\left| P_{i|V}(x) - \hat{P}_{i|V}(x) \right| \leq \frac{\left| \hat{P}(x(V)) - P(x(V)) \right| + \hat{P}_{i|V}(x) \left| P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right|}{P(x(V/\{i\}))}.$$

Hence,

$$\left| P_{i|V}(x) - \hat{P}_{i|V}(x) \right| \leq 2\sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V/\{i\}))}} \left( \sqrt{P_{i|V}(x)} + \hat{P}_{i|V}(x) \right).$$

We just prove that

$$\frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \leq \left( \frac{1 + 2\Lambda^{-1/2}}{1 - 2\Lambda^{-1/2}} \right)^2, \text{ hence } \hat{P}_{i|V}(x) \leq \sqrt{\hat{P}_{i|V}(x)} \leq \frac{1 + 2\Lambda^{-1/2}}{1 - 2\Lambda^{-1/2}} \sqrt{P_{i|V}(x)}.$$

Therefore,

$$\begin{aligned} \left| P_{i|V}(x) - \hat{P}_{i|V}(x) \right| &\leq \frac{4}{1 - 2\Lambda^{-1/2}} \sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V/\{i\}))}} \sqrt{P_{i|V}(x)} \\ &\leq \frac{4}{1 - 2\Lambda^{-1/2}} \sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V))}} P_{i|V}(x) \leq \frac{4}{\Lambda^{1/2} - 2} P_{i|V}(x). \end{aligned}$$

Let  $u^2 = n^{-1} \ln(2a^s N_s \delta)$ . On  $\Omega_{prob}$ , we have

$$\widehat{P}(x(V)) \leq P(x(V)) + 2\frac{u}{\sqrt{2}}\sqrt{P(x(V))} + \frac{u^2}{3} = \left(\sqrt{P(x(V))} + \frac{u}{\sqrt{2}}\right)^2 + \frac{u^2}{12}.$$

Since  $\widehat{P}(x(V)) \geq \Lambda u^2$ , we deduce that

$$P(x(V)) \geq \left(\sqrt{\Lambda - \frac{1}{12}} - \frac{1}{\sqrt{2}}\right)^2 u^2 = \left(\Lambda + \frac{5}{12} - \sqrt{\frac{6\Lambda - 1}{3}}\right) u^2.$$

Since  $\Lambda \geq 2$ , we deduce that

$$\left|\widehat{P}(x(V)) - P(x(V))\right| \leq \left(\sqrt{2} + \frac{1}{3\sqrt{\Lambda + \frac{5}{12} - \sqrt{\frac{6\Lambda - 1}{3}}}}\right) u\sqrt{P(x(V))} \leq 2u\sqrt{P(x(V))}.$$

From the same inequality, we also obtain

$$\left|\widehat{P}(x(V)) - P(x(V))\right| \leq \frac{2P(x(V))}{\sqrt{\Lambda - \frac{1}{12}} - \frac{1}{\sqrt{2}}} \leq \frac{2P(x(V))}{\sqrt{\Lambda}}.$$

Since  $\widehat{P}(x(V/\{i\})) \geq \widehat{P}(x(V))$ , we prove with the same arguments that

$$\left|P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right| \leq \sqrt{\frac{P(x(V/\{i\})) \ln(2a^s N_s \delta)}{n}} \leq \frac{2P(x(V/\{i\}))}{\sqrt{\Lambda}}.$$

From Lemma 8.1, we have

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \widehat{P}_{i|V}(x) \left|P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right|}{P(x(V/\{i\}))}.$$

Hence,

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 2\sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V/\{i\}))}} \left(\sqrt{P_{i|V}(x)} + \widehat{P}_{i|V}(x)\right).$$

If  $\sqrt{P_{i|V}} \geq \widehat{P}_{i|V}$ , we deduce that

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 4\sqrt{\frac{\ln(2a^s N_s \delta)P_{i|V}(x)}{nP(x(V/\{i\}))}}.$$

Otherwise, we have

$$\left(1 - \frac{4}{\sqrt{\Lambda}}\sqrt{1 + \frac{2}{\sqrt{\Lambda}}}\right) \widehat{P}_{i|V}(x) \leq \left(1 - 4\sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V/\{i\}))}}\right) \widehat{P}_{i|V}(x) \leq P_{i|V}(x).$$

Since  $\Lambda \geq 100$ , we obtain  $\widehat{P}_{i|V}(x) \leq 2P_{i|V}(x)$ . We deduce that

$$\begin{aligned} \left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| &\leq 2(1 + \sqrt{2})\sqrt{\frac{\ln(2a^s N_s \delta)P_{i|V}(x)}{nP(x(V/\{i\}))}} = 2(1 + \sqrt{2})\sqrt{\frac{\ln(2a^s N_s \delta)}{nP(x(V))}}P_{i|V}(x) \\ &\leq \frac{2(1 + \sqrt{2})}{\sqrt{\Lambda}}\sqrt{1 + \frac{2}{\sqrt{\Lambda}}}P_{i|V}(x) \leq \sqrt{\frac{11}{\Lambda}}P_{i|V}(x). \end{aligned}$$

## 8.6 Control of the variance terms in Küllback loss:

The following Lemma gives an important decomposition of the Küllback loss.

**Lemma 8.21.** *Let  $\Lambda \geq 100$  and let  $\mathcal{V}_{s,\Lambda}$ ,  $\mathcal{V}_{s,\Lambda}^{(2)}$  be respectively the collection of subsets  $V$  in  $\mathcal{V}_s$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,*

$$P(x(V) = 0, \text{ or } \hat{P}(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n}$$

*and the collection of subsets  $V$  in  $\mathcal{V}_s$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,*

$$P(x(V) = 0, \text{ or } P(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n}.$$

*Let  $\Omega_{\text{prob}}(\delta)$  be the event defined on Lemma 8.20. On  $\Omega_{\text{prob}}(\delta)$ , for all  $V$  in  $\mathcal{V}_{s,\Lambda}$ , we have*

$$p_1(V) \leq \frac{20}{6} \sum_{x \in \mathcal{X}(V)} \frac{\left( P(x(V)) - \hat{P}(x(V)) \right)^2}{P(x(V))} + \frac{14}{9} \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left( P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))}.$$

$$p_2(V) \leq \frac{3}{2} \sum_{x \in \mathcal{X}(V)} \frac{\left( P(x(V)) - \hat{P}(x(V)) \right)^2}{P(x(V))} + \frac{7}{3} \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left( P(x(V/\{i\})) - \hat{P}(x(V/\{i\})) \right)^2}{P(x(V/\{i\}))}.$$

**Proof:** From Lemma 8.20, for all  $V$  in  $\mathcal{V}_{s,\Lambda}$  or  $\mathcal{V}_{s,\Lambda}^{(2)}$ , for all  $x$  in  $V$ , we have  $|P_{i|V}(x), \hat{P}_{i|V}(x)| \leq \sqrt{11\Lambda^{-1}} P_{i|V}(x)$ . Hence, from Lemma 8.18,

$$\begin{aligned} p_1(V) &= \sum_{x \in \mathcal{X}(V)} P(x(V)) \ln \left( \frac{P_{i|V}(x)}{\hat{P}_{i|V}(x)} \right) \\ &\leq \frac{1}{2} \left( 1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} P(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{\hat{P}_{i|V}(x)}. \\ p_2(V) &= \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right) \\ &\leq \frac{1}{2} \left( 1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{P_{i|V}(x)}. \end{aligned}$$

From Lemma 8.1, we have

$$\begin{aligned} |P_{i|V}(x) - \hat{P}_{i|V}(x)| &\leq \frac{|P(x(V)) - \hat{P}(x(V))| + P_{i|V}(x) |P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))|}{\hat{P}(x(V/\{i\}))} \\ |P_{i|V}(x) - \hat{P}_{i|V}(x)| &\leq \frac{|P(x(V)) - \hat{P}(x(V))| + \hat{P}_{i|V}(x) |P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))|}{P(x(V/\{i\}))}. \end{aligned}$$

The second inequality gives that  $p_1(V)$  is smaller than

$$\begin{aligned} &\leq \left(1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right) \sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \hat{P}(x(V))\right)^2}{P(x(V/\{i\}))\hat{P}_{i|V}(x)} + \hat{P}_{i|V}(x) \frac{\left(P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))} \\ &\leq \frac{1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}}}{1 - \sqrt{\frac{11}{\Lambda}}} \sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \hat{P}(x(V))\right)^2}{P(x(V))} + \left(1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right) \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left(P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}. \end{aligned}$$

We also have

$$\begin{aligned} \left(P_{i|V}(x) - \hat{P}_{i|V}(x)\right)^2 &\leq \frac{\left(P(x(V)) - \hat{P}(x(V))\right)^2 + \hat{P}_{i|V}(x)P_{i|V}(x)P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))^2}{P(x(V/\{i\}))\hat{P}(x(V/\{i\}))} \\ &+ \frac{(\hat{P}_{i|V}(x) + P_{i|V}(x)) \left|P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right| \left|P(x(V)) - \hat{P}(x(V))\right|}{P(x(V/\{i\}))\hat{P}(x(V/\{i\}))} \\ &\leq \frac{3 \left(P(x(V)) - \hat{P}(x(V))\right)^2 + 2(\hat{P}_{i|V}(x) + P_{i|V}(x))^2 \left(P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right)^2}{2P(x(V/\{i\}))\hat{P}(x(V/\{i\}))}. \end{aligned}$$

Hence,

$$\begin{aligned} p_2(V) &\leq \frac{3}{2} \sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \hat{P}(x(V))\right)^2}{P(x(V))} \\ &+ \sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))} \frac{(\hat{P}_{i|V}(x) + P_{i|V}(x))^2}{P_{i|V}(x)} \end{aligned}$$

is smaller than

$$\frac{3}{2} \sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \hat{P}(x(V))\right)^2}{P(x(V))} + \left(2 + \sqrt{\frac{11}{\Lambda}}\right) \sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left(P(x(V/\{i\})) - \hat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

## 8.7 Concentration for the slope heuristic in the Küllback case

**Lemma 8.22.** *Let  $\Lambda \geq 100$  and let  $\mathcal{V}_{s,\Lambda}$  and  $\mathcal{V}_{s,\Lambda}^{(2)}$  be respectively the collection of subsets  $V$  in  $\mathcal{V}_s$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,*

$$P(x(V)) = 0, \text{ or } \hat{P}(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n}$$

*and the collection of subsets  $V$  in  $\mathcal{V}_s$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,*

$$P(x(V)) = 0, \text{ or } P(x(V)) \geq \Lambda \frac{\ln(2a^s N_s \delta)}{n}.$$

*Let  $\Omega_{\text{prob}}(\delta)$  be the event defined on Lemma 8.20. On  $\Omega_{\text{prob}}(\delta)$ , there exists an absolute constant  $C > 0$  such that*

$$|p_1(V) - p_2(V)| \leq \frac{C}{\sqrt{\Lambda}} p_1(V).$$

**Proof:** We use Lemmas 8.18 and 8.20. On  $\Omega_{\text{prob}}(\delta)$ , we have

$$\begin{aligned}
& \frac{1}{2} \left( 1 - \frac{7\sqrt{11}}{3\sqrt{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} P(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{\hat{P}_{i|V}(x)} \leq p_1(V). \\
p_1(V) & \leq \frac{1}{2} \left( 1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} P(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{\hat{P}_{i|V}(x)}. \\
& \frac{1}{2} \left( 1 - \frac{7\sqrt{11}}{3\sqrt{\Lambda}} \right) \left( 1 - \sqrt{\frac{11}{\Lambda}} \right) \left( 1 - \sqrt{\frac{4}{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} P(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{\hat{P}_{i|V}(x)} \\
& \leq p_2(V) = \sum_{x \in \mathcal{X}(V)} \hat{P}(x(V)) \ln \left( \frac{\hat{P}_{i|V}(x)}{P_{i|V}(x)} \right) \\
& \leq \frac{1}{2} \left( 1 + \frac{5\sqrt{11}}{3\sqrt{\Lambda}} \right) \left( 1 + \sqrt{\frac{11}{\Lambda}} \right) \left( 1 + \sqrt{\frac{4}{\Lambda}} \right) \sum_{x \in \mathcal{X}(V)} P(x(V/\{i\})) \frac{(P_{i|V}(x) - \hat{P}_{i|V}(x))^2}{\hat{P}_{i|V}(x)}.
\end{aligned}$$

## 8.8 Concentration of $\mathbf{L}(V)$ - $\mathbf{L}(V')$

The following Lemma let us control the remainder term in the oracle inequality.

**Lemma 8.23.** *Let  $\delta > 1$  and let  $\mathcal{V}_{s,\Lambda,p_*}$  be the subset of  $\mathcal{V}_{s,\Lambda}$  of the sets  $V$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,  $P_{i|V}(x) = 0$  or  $\hat{P}_{i|V}(x) \geq p_*$ . With probability at least  $1 - \delta$ , for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}$ , for all  $\eta > 0$ , we have,*

$$\begin{aligned}
& \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \ln \left( \frac{P_{i|V}(x)}{P_{i|V'}(x)} \right) \\
& \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).
\end{aligned}$$

*Let  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$  be the subset of  $\mathcal{V}_{s,\Lambda}^{(2)}$  of the sets  $V$  such that, for all  $x$  in  $\mathcal{X}(V)$ ,  $P_{i|V}(x) = 0$  or  $P_{i|V}(x) \geq p_*$ . With probability at least  $1 - \delta$ , for all  $V, V'$  in  $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$ , for all  $\eta > 0$ , we have,*

$$\begin{aligned}
& \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \ln \left( \frac{P_{i|V}(x)}{P_{i|V'}(x)} \right) \\
& \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\ln(N_s^2 \delta)}{n} \left( 4 \ln n + \frac{3}{2\eta p_*} \right).
\end{aligned}$$

**Proof:** Let us first write

$$\begin{aligned}
& \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \ln \left( \frac{P_{i|V}(x)}{P_{i|V'}(x)} \right) \\
& \leq \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V'))) \left( \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V'}(x)} \right) - \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V}(x)} \right) \right).
\end{aligned}$$

Let us now write  $V_*$  for  $V$  or  $V'$ . We have

$$\begin{aligned} \sum_{x \in \mathcal{X}(V \cup V')} (\hat{P}(x(V \cup V')) - P(x(V \cup V')) \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V_*}(x)} \right)) \\ = (P_n - P) \left( \sum_{x \in \mathcal{X}(V \cup V')} \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V_*}(x)} \right) \mathbf{1}_{x(V \cup V')} \right). \end{aligned}$$

The function  $f : \mathcal{X}(V \cup V') \rightarrow \mathbb{R}$ ,  $x \mapsto \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V_*}(x)} \right)$  is upper bounded on  $\Omega_{prob}(\delta)$  by  $2 \ln n$ . Since it is not random, the bound also holds on  $\Omega_{prob}(\delta)^c$ . Let us evaluate its variance

$$\text{Var}(f(X)) \leq P f^2 = \sum_{x \in \mathcal{X}(V \cup V')} P(x(V \cup V')) \left( \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V_*}(x)} \right) \right)^2$$

Let us recall also here the following Lemma see [Mas07], Lemma 7.24 p 275 or [BS91])

**Lemma 8.24.** *For all probability measures  $P$  and  $Q$ , with  $P \ll Q$ ,*

$$\frac{1}{2} \int (dP \wedge dQ) \left( \ln \left( \frac{dP}{dQ} \right) \right)^2 \leq K(P, Q) \leq \frac{1}{2} \int (dP \vee dQ) \left( \ln \left( \frac{dP}{dQ} \right) \right)^2.$$

Since  $P_{i|V_*}(x) \geq 2\hat{P}_{i|V_*}(x)/3 \geq 2p_*/3$ , we deduce that

$$\text{Var}(f(X)) \leq \frac{3}{p_*} K(P_{i|V \cup V'}, P_{i|V_*}).$$

Applying Benett's inequality to  $f$ , we obtain that, with probability  $1 - 2e^{-t}$ ,

$$(P_n - P) \left( \sum_{x \in \mathcal{X}(V \cup V')} \ln \left( \frac{P_{i|V \cup V'}(x)}{P_{i|V_*}(x)} \right) \mathbf{1}_{x(V \cup V')} \right) \leq \sqrt{\frac{6}{p_*} K(P_{i|V \cup V'}, P_{i|V_*}) \frac{t}{n}} + \frac{2t \ln n}{n}.$$

We conclude the proof with a union bound and the classical inequality  $2ab \leq \eta a^2 + \eta^{-1} b^2$ .

## References

- [AB10] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, 22:46–54, 2010.
- [AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, 10:245–279, 2009.
- [Bes93] L. Besag. Statistical analysis of dirty pictures. *Journal of applied statistics*, 20(5):63–87, 1993.
- [BKM04] E.N. Brown, R.E. Kass, and P.P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7:456–461, 2004.
- [BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

- [BM09] J. Bento and A. Montanari. Which graphical models are difficult to learn? *available Markov random field texture models* at <http://arxiv.org/pdf/0910.5761>, 2009.
- [BMS08] G. Bresler, E. Mossel, and A. Sly. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, chapter Reconstruction of Markov Random Fields from Samples: Some Easy Observations and Algorithms, pages 343–356. Springer, 2008.
- [Bou02] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.
- [BS91] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.
- [CJ83] G. Cross and A. Jain. Markov random field texture models. *IEEE Trans. PAMI*, 5:25–39, 1983.
- [CT06a] I. Csiszar and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. *Ann. Statist.*, 34:123–145, 2006.
- [CT06b] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3):1007–1016, 2006.
- [FP97] R. Fernández and C.-E. Pfister. Global specifications and nonquasilocality of projections of gibbs measures. *The Annals of Probability*, 25:1284–1315, 1997.
- [GOT10] A. Galves, E. Orlandi, and D. Y. Takahashi. Identifying interacting pairs of sites in infinite range ising models. *Preprint*, <http://arxiv.org/abs/1006.0272>, 2010.
- [KT63] A. Kolmogorov and V. Tikhomirov.  $\epsilon$ -entropy and  $\epsilon$ -capacity of sets in functional spaces. *Amer. Math. Soc. Trans.*, 1(7):277–364, 1963.
- [Ler11a] M. Lerasle. Optimal model selection for stationary data under various mixing conditions. *to appear in Ann. Statist.*, 2011.
- [Ler11b] M. Lerasle. Optimal model selection in density estimation. *to appear in Ann. I. H. Poincaré*, 2011.
- [LOU<sup>+</sup>10] X. Li, G. Ouyang, A. Usami, Y. Ikegaya, and A. Sik. Scale-free topology of the ca3 hippocampal network: a novel method to analyze functional neuronal assemblies. *Biophysics Journal*, 98:1733–1741, 2010.
- [LT11] M. Lerasle and D. Yasumasa Takahashi. An oracle approach for interaction neighborhood estimation in random fields. *to appear in Electron. J. Statist. arXiv:1010.4783*, 2011.
- [Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- [Rip81] B.D. Ripley. *Spatial Statistics*. Wiley, New York, 1981.
- [RWL10] P. Ravikumar, M.J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using  $l_1$ -regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

- [Sau11] A. Saumard. Nonasymptotic quasi-optimality of aic and the slope heuristics in maximum likelihood estimation of density using histogram models. *hal-00512310*, 2011.
- [SBSB06] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.
- [TSMI10] N. Takahashi, T. Sasaki, W. Matsumoto, and Y. Ikegaya. Circuit topology for synchronizing neurons in spontaneously active networks. *Proceedings of National Academy of Science U.S.A.*, 107:10244–10249, 2010.
- [Ver09] N. Verzelen. Adaptive estimation of stationary gaussian fields. Technical Report RR-6797, INRIA, 2009.
- [Woo78] J. Woods. Markov image modeling. *IEEE Trans. Automat. Control*, 23:846–850, 1978.
- [Yan05] Y. Yang. Can the strength of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.